University of Alberta

Estimation of Principal Points using Smoothing

by

Mawuli Foli Kuivi

A thesis

submitted to the Faculty of Graduate Studies and Research in partial

fulfillment of the requirements for the degree of Master of Science

in

Statistics

Department of Mathematical Sciences

Edmonton, Alberta

Fall 1997

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-22619-0

Canada

University of Alberta

Library Release Form

Name of Author: Mawuli Foli Kuivi

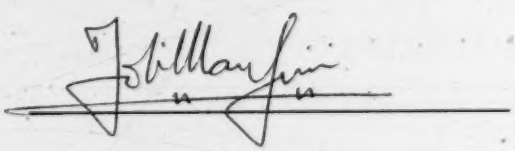
Title of Thesis: Estimation of Principal Points using Smoothing

Degree: Master of Science

Year this Degree Granted: 1997

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided, neither the thesis nor any substantial portion thereof may may be printed or otherwise reproduced in any material from whatever without the author's prior written permission.



Permanent Address:

P . O . Box 404, HO

Ghana

West Africa.

Date: 27 June, 1997

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled Estimation of Principal Points using Smoothing submitted by Mawuli Foli Kuivi in partial fulfillment of the requirements for the degree of Master of Science in Statistics.

Peter Hooper

P. M. Hooper (Supervisor)

K C Carrière

K. C. Carrière

W. W. Armstrong

W. W. Armstrong (External Examiner)

Date: June 25, 1997

TO MY

DAD, MUM, and BROTHERS

ABSTRACT

A set of K principal points for the distribution of a random vector X is a set of K points minimizing the expected squared distance between X and the nearest point in the set. The thesis is concerned with the use of smoothing in the estimation of principal points.

Problems considered include the estimation of the risk function and the estimation of the optimal smoothing parameter τ_{PP} when estimating principal points.

The thesis commences with a review of some literature on principal points and related topics.

The second chapter investigates the effectiveness of optimal smoothing in the context of the spherical normal distribution.

The third chapter treats the estimation of an optimal smoothing parameter and evaluation of its effectiveness given a data set from an unknown distribution.

These techniques are applied to several examples in Chapter 4.

ACKNOWLEDGEMENTS

I wish to thank my thesis supervisor, Dr. P. M. Hooper, for his guidance through my new field. I am very grateful for his continued readiness to listen and help. It is very much appreciated.

I would also like to thank Dr. W. W. Armstrong and Dr. K. C. Carrière for reading my thesis and for their suggestions. I also express my eternal gratitude to the staff of the Department of Mathematical Sciences which, with so much generosity, helped me make achievements I could not have done by myself.

I also wish to express my appreciation to the University of Alberta and the Natural Sciences and Engineering Research Council of Canada for their financial support.

Lastly, but not the least, thanks to the Lord for keeping me safe and my family for their wonderful faith in me.

Contents

1 Introduction And Background **1**

1.1 Definitions and Examples 1

1.2 K–Means Clustering Algorithm. 4

1.3 Stochastic Approximation 7

1.4 Vector Quantization 8

1.5 Analytical Results 10

1.5.1 Self–Consistent Points 10

1.5.2 Principal Points 11

1.5.3 Principal Components 13

2 Smoothing **15**

2.1 Definitions and Rationale 15

2.2 Optimal Smoothing for the Spherical Normal 17

3 Estimation of Principal Points **28**

3.1 Introduction 28

3.2 v–Fold Cross–Validation 28

3.3 Optimal Smoothing Estimated from a Data Set 30

4 Numerical Examples **33**

4.1 Representative Curves 33

4.2 Ozone Data 35

4.3 Boston Housing Data 36

4.4 Seals Data 39

4.5 Conclusion 41

Bibliography **57**

List of Tables

2.1 Table for Design of Model. 24

2.2 Table for the results for the model approximation. 24

4.1 Results for Representative Curves. 42

4.2 Results for Ozone Data. 46

4.3 Results for Boston Housing Data. 50

4.4 Results for Seals Data. 53

List of Figures

2.1 (a): Plot of residuals versus number of principal points, K and (b): Plot of τ_{PP}/τ_{DE} versus the number of principal points, K. . . 25

2.2 Plot of τ_{PP}/τ_{DE} versus the dimension, p. 26

2.3 Plot of R_{opt}/R_0 versus linear combination of n, p, and K as given in (2.19). 27

4.1 A collection of Density Estimates (Example 1). One hundred kernel density estimates were used based on independent samples of size fifty from the standard normal distribution overlaid on one another 43

4.2 Representative Curves based on ($K = 2$ and 3) principal point estimates in Example 1 (Density Estimation). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines). 44

4.3 Representative Curves based on ($K = 4$ and 5) principal point estimates in Example 1 (Density Estimation). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines). 45

4.4 The Ozone Data (Example 2). Plot of ozone data by linearly interpolating values for each week. On the x-axis, 1 = Monday,...,7 = Sunday and ozone levels on the y-axis. 47

4.5 Representative Curves based on ($K = 2$ and 3) principal point estimates in Example 2 (Ozone data). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines). On the x-axis, 1 = Monday,...,7 = Sunday and ozone levels on the y-axis. 48

4.6 Representative Curves based on ($K = 4$ and 5) principal point estimates in Example 2 (Ozone data). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines). On the x-axis, 1 = Monday,...,7 = Sunday and ozone levels on the y-axis. 49

4.7 Representative Curves based on ($K = 2$ and 3) principal point estimates in Example 3 (Boston Housing Data). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines). 51

4.8 Representative Curves based on ($K = 4$ and 5) principal point estimates in Example 3 (Boston Housing Data). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines). 52

4.9 Representative Curves based on ($K = 2$ and 3) principal point estimates in Example 4 (Seals data). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines). 54

4.10 Representative Curves based on principal point estimates in Example 4 (Seals data). The curves based on ($K = 4$ and 5) principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines). 55

4.11 Representative Curves for each of the herds based on ($K = 1$ and 2) principal point estimates in Example 4 (Seals data). 56

Chapter 1

Introduction And Background

1.1 Definitions and Examples

The K principal points of a p–dimensional random vector X are those points $\xi_1, \xi_2, \ldots, \xi_K \in \Re^p$ that minimize the expected square distance of X from the nearest of the ξ_k (Flury, 1990).

More formally, define $K_* : \Re^{p\times(1+K)} \longrightarrow \{1, 2, \ldots, K\}$ by

$$K_*(x, y_1, \ldots, y_K) = \min\{k : \|x - y_k\| \leq \|x - y_l\| \text{ for } 1 \leq l \leq K\}. \qquad (1.1)$$

The expression K_* is the index of the point y_k that is closest to the point x. If x has equal distance to two or more y_k, then K_* is the first such index in the list.

Define

$$\delta_k(x) = \begin{cases} 1, & \text{if } \quad k = K_*(x, y_1, \ldots, y_K) \\ 0, & \text{otherwise .} \end{cases}$$

The dependence of $\delta_k(x)$ on $y_1, y_2, \ldots, y_K$ is implicit in this notation. Sometimes we will make the dependence explicit by writing $\delta_k(x, y_1, y_2, \ldots, y_K)$.

Let X be p–dimensional random vector with distribution function F. Define

the loss function $L_F : \Re^{p \times K} \longrightarrow \Re$ by

$$L_F(y_1, y_2, \ldots, y_K) = E_F \left\{ \sum_{k=1}^{K} \delta_k(X) \|X - y_k\|^2 \right\}, \tag{1.2}$$

where $\|x - y\|^2 = (x - y)^T (x - y)$. A set of points minimizing (1.2) is called a set of K principal points for F and is denoted by $\xi_1, \xi_2, \ldots, \xi_K$. If F is replaced by $\hat{F}$, the empirical distribution function of a sample $\{x_1, x_2, \ldots, x_n\}$, the loss function (1.2) becomes

$$L_{\hat{F}}(y_1, y_2, \ldots, y_K) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \le k \le K} \|x_i - y_k\|^2. \tag{1.3}$$

When an estimate such as $\hat{F}$ is used, then we will denote the estimated principal points by $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K$.

The term principal points was introduced by Flury (1990) as a result of a practical problem in statistical consulting. It started with a project of the Swiss Army which wanted to design new protection masks. To put the construction of the new masks on a good empirical grounds, a group of anthropologists was hired to measure the heads of 900 Swiss soldiers. Twenty five variables were regarded as potentially important for the fit of the masks. Since human heads differ in size and shape, several types of masks are needed for adequate fit. Flury considered how one might determine an optimal set of K types, where $K \le 5$. The optimality criterion (1.3) was used to construct models of K typical heads, which were then used to assist in designing masks.

The solution to the mask fitting problem is computationally equivalent to the K-means clustering algorithm, which attempts to find optimal cluster means for a multivariate sample using criterion (1.3). Since cluster analysis is usually related to the idea of finding homogeneous subgroups in a mixture of distributions,

the name principal points was suggested for optimal cluster means in a homogeneous population. In addition, the term principal points was introduced as a way of emphasizing a whole range of applications beyond clustering and a connection with principal components.

Principal points were studied in the theory of stratified sampling by Dalenius (1950); Dalenius and Gurvey (1951) and Cox (1957). These authors studied how to partition a given set of data or items into K homogeneous subsamples. Their work was restricted to univariate distributions.

Flury and Tarpey (1993) used principal points in the selection of representative curves from a collection of curves. Suppose we have N curves. Approximate each curve by a p–dimensional vector obtained by evaluating the curves at equispaced points. Compute a set of K principal points for the p–dimensional data set, plot the coordinates, and interpolate to obtain K smooth curves.

The K–means clustering criterion (1.3) was introduced by MacQueen (1965). He suggested applications in classification as well as clustering, i.e., for each class, find a set of K principal points and use these for nearest neighbor classification.

Finding principal points can be seen as data reduction. In the Signal Processing literature, this technique is called Vector Quantization. In the Information Theory literature, it is called Source Coding with Fidelity Criterion. The data set may consist of blocks of pixels extracted from a training image. The dimension of this data set is reduced using vector quantization (described later in section 1.4) and the reduced version of the data set is either transmitted or stored (Cohn et al, 1994).

Principal point estimates have been used as an initial step in two classification algorithms; learning vector quantization (Kohonen, 1995) and piecewise

linear classification (Hooper, 1996), and in two regression algorithms, radial basis functions regression (Moody and Darken, 1989) and normalized exponential smoothing (Hooper, 1996).

The thesis is organized as follows. We begin with the review of some literature in chapter one. In chapter two, we introduce the concept of smoothing. Here we look at smoothing when estimating principal points for a particular distribution. Chapter three also looks at smoothing but rather for a general data set. Also this chapter deals with estimating the loss function for a data set. Vector quantization is applied to estimate principal points in both chapters two and three. Some numerical examples are given in chapter four which is the last chapter.

1.2 K–Means Clustering Algorithm.

Partitioning methods in cluster analysis are usually based on an optimization criterion that measures compatibility of clustering parameters with a data set describing the objects. In general, an optimal solution cannot be obtained in a closed form and iterative algorithms are necessary.

The K–means clustering algorithm (Hartigan and Wong, 1979) is a special case of the Classification EM algorithm (Celeux and Govaert, 1992). Let $x_1, x_2, \ldots, x_n \in \Re^p$ be a sample from a mixture of densities

$$f(x) = \sum_{k=1}^{K} p_k f(x, \mathbf{a}_k), \tag{1.4}$$

where the p_k's are the mixing weights, $0 < p_k < 1$, for all $k = 1, 2, \ldots, K$ and $\sum_k p_k = 1$, and the $f(x, \mathbf{a}_k)$ are densities from the same parametric family. For example, $f(x, \mathbf{a}_k)$ might denote the p–dimensional normal density with unknown mean μ_k and covariance matrix Σ, and $\mathbf{a}_k = (\mu_k, \Sigma)$.

The EM algorithm is a general algorithm to compute the maximum likelihood estimates of p_k, $\mathbf{a}_k, 1 \leq k \leq K$ under the mixture approach. The Classification EM (CEM) algorithm is a general algorithm to compute the estimates $\mathbf{a}_k, p_k$ and to find the clusters $\pi_k, 1 \leq k \leq K$ under the classification approach. The CEM algorithm incorporates the E–step (Expectation) and the M-step (Maximum likelihood estimation) of the EM algorithm using a maximum posterior principle.

The CEM is described as follows. Let $\pi^m = \{\pi_1{}^m, \pi_2{}^m, \ldots, \pi_K{}^m\}$ denote partition of sample points $\{1, 2, \ldots, n\}$ into K subsets. Start with an initial partition π^0. The mth iteration of the CEM algorithm is given as:

E–Step:- Compute for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, K$ the current posterior probabilities $t_k{}^m(x_i)$ that x_i belongs to π_k as

$$t_k{}^m(x_i) = \frac{p_k^m f(x_i, \mathbf{a}_k^m)}{\sum_{k'=1}^{K} p_{k'}^m f(x_i, \mathbf{a}_{k'}^m)} \tag{1.5}$$

for the current parameter estimates $\mathbf{p}^m$ and $\mathbf{a}^m$.

C–Step:- Assign each x_i to the cluster which provides the maximum probability $t_k{}^m(x_i), k = 1, 2, \ldots, K$. If the maximum posterior probability is not unique, we assign to the cluster with the smallest index. Let π^m denote the resulting partition.

M–Step:- For $k = 1, 2, \ldots, K$, compute the maximum likelihood estimates $p_k{}^{m+1}$, $\mathbf{a}_k^{m+1}$ using the subsamples $\pi_k{}^m$. We have

$$p_k{}^{m+1} = \frac{\#\pi_k{}^m}{n}, \quad \text{for all} \quad k = 1, 2, \ldots, K. \tag{1.6}$$

The formula for the $\mathbf{a}_k^{m+1}$'s depends on the family of density functions involved.

The K–means clustering algorithm can be obtained as a special case of the CEM algorithm assuming a Gaussian mixture with equal proportions and common covariance matrix of the form $\sigma^2 I$. The estimation of the scale parameter σ^2 does not affect the assignment of the x_i's to the clusters π_k^m. In the M–step we have

$$\mu_k^{m+1} = \frac{1}{\#\pi_k^m} \sum_{x_i \in \pi_k^m} x_i, \quad \text{for all} \quad k = 1, 2, \ldots, K\,, \tag{1.7}$$

and

$$\left(\sigma^2\right)^{m+1} = \frac{1}{np} \sum_{k=1}^{K} \sum_{x_i \in \pi_k^m} \left\| x_i - \mu_k^{m+1} \right\|^2. \tag{1.8}$$

In the C–step, x_i is assigned to the cluster with nearest centroid mean $\mu_k^{(m)}$.

One can specify an initial partition or initial centroids. The final partition will be, to some extent, dependent upon the initial partition or the initial specified centroids. It is therefore advisable to repeat the clustering using different initial partitions or initial centroids.

A stochastic version of the K–means algorithm was developed by Celeux and Govaert (1992) which improves the chance of finding a global minimum. In its second step (classification step), the algorithm determines, for each x_i, a probability distribution over the set of all K cluster means and then assigns x_i randomly according to this distribution. The probabilities are similar to posterior probabilities of cluster membership calculated assuming a normal mixture model. The probabilities are gradually adjusted so that, as the number of iterations increases, the probability that x_i is assigned to the nearest cluster mean tends to one.

1.3 Stochastic Approximation

In many optimization problems, the solution is a vector of parameter values that minimizes a given performance index or objective function, usually expressed as an integral. There are problems where such an optimum can be determined in closed form but often optimization is not analytically tractable. Robbins and Monro (1951) introduced an iterative technique for optimization called stochastic approximation. Benveniste et al. (1990) give a general review of theory and applications of stochastic approximation. A brief summary is as follows. We wish to minimize a function $R(\theta)$ by using an iterative algorithm driven by a sequence of independent and identically distributed random vectors $\{\mathbf{Z}_m\}$:

$$\theta_m = \theta_{m-1} + a_m H(\theta_{m-1}, \mathbf{Z}_m). \tag{1.9}$$

The term a_m is called the gain function and H is the updating function. Let the gain function satisfy ;

$$a_m > 0, \ \sum a_m = \infty, \quad \text{and} \quad \sum a_m^{\alpha} < \infty \quad \text{for some } \alpha > 1. \tag{1.10}$$

Write

$$\theta_m = \theta(t_m) \quad \text{where} \quad t_m = \sum_{i=1}^{m} a_i. \tag{1.11}$$

After an initial transient phase, the behavior of algorithm (1.9) is represented to a first approximation by the ordinary differential equation

$$\frac{d}{dt}\theta(t) = h(\theta(t)), \tag{1.12}$$

where

$$h(\theta) = E\{H(\theta, \mathbf{Z})\}. \tag{1.13}$$

In gradient algorithms, the updating function H is defined so that $h(\theta) = -\bigtriangledown R(\theta)$.

1.4 Vector Quantization

Vector quantization is a classical method in signal processing that approximates a distribution by a representative set of vectors (Gersho and Gray, 1992). One optimal approximation produces a set of K vectors y_k so that the expected squared distance from the nearest y_k is minimized.

As mentioned earlier, the data set (vectors x) may consist of blocks of pixels extracted from a training image. The image is broken up into rectangular pixel blocks. Each of these blocks is a p–dimensional vector. This p is equal to the number of pixels within a block. The quantity which is being measured is the light intensity for each pixel. Hence to obtain a vector, we take the value of the light intensity of each pixel in a block.

A training set of ten 512×512 pixel images can be broken into blocks of $4 \times 4 = 16$ pixels. So the dimension in this case is 16 and there are $128 \times 128 = 16,384$ blocks per image. This gives 163,840 blocks for a training set. In other words, we have a data set of 163,840 vectors each of dimension 16. A set of principal points is estimated using vector quantization. This reduced set of vectors is called a codebook and it is used to encode the image. The reduced set of vectors is either stored or transmitted instead of the entire data set. This reduces the transmission costs.

Encoding a gray–scale image (with 8 bits per pixel or 2^8 different levels of light intensity per pixel) with a codebook of 256 4×4 blocks will require that $\log_2(256) = 8$ bits be used for every $8 \times 4 \times 4 = 128$–bit block, resulting in a 16–to–1 compression ratio. This compression will affect the quality of the final image to some extent.

Let $\hat{F}$ be the estimate of the distribution function of the vectors in the image. Also, let $\{X_m\}$ be a sequence of independent random vectors sampled from $\hat{F}$. Then the function we would like to minimize is

$$E_{\hat{F}}\left\{\sum_{k=1}^{K}\delta_k(X)\|X-y_k\|^2\right\}.$$

Vector quantization may then be implemented by stochastic approximation with updating formula

$$Y_k \longleftarrow Y_k + a_m\delta_k(X_m)(X_m - Y_k). \tag{1.14}$$

This means that at the mth iteration the point Y_k closest to X_m is moved towards X_m.

Let M be the number of iterations chosen to evaluate the loss function. We define the gain function a_m so that $1/a_m$ varies linearly from $1/a_1$ to $1/a_M$. The rate of convergence is determined by the values of a_1, a_M and M. Decreasing the gain slowly is likely to give convergence to a point near optimum. For example when cooling metals, if the metal is cooled slowly, the resulting crystal is at a state of minimum energy but when cooled quickly it does not reach this state, but rather a state of higher energy. Vector quantization can be used to approximate a set of principal points for smoothed distribution function estimates.

It appears that the algorithm spends most of its time calculating distances to obtain $K_*(x, y_1, y_2, \ldots, y_K)$ so as to obtain $\delta_k(x)$. The computing time, t_{comp} required to implement the algorithm depends primarily on K and p, but not on n. A double-precision FORTRAN implementation of the algorithm was run on RS/6000 Model 350 workstation using values of K ranging from 2 to 20, p ranging from 1 to 10 and for a fixed value of M = 100,000. The computing

times, t_{comp} for this particular M varied from 12 to 74 seconds and are closely approximated ($R^2 = 99.35\%$) by the formula

$$t_{comp} = 7.6782 + 0.4243K + 3.7528p + 0.0915Kp. \tag{1.15}$$

The algorithm is similar to that of MacQueen (1965) described as follows. Start with K clusters each consisting of a single random point from the set. Next, select a point randomly from the entire data set and add it to the cluster whose mean the new point is nearest. After adding a new point to a cluster, the mean of that cluster is recalculated in order to take account of the new point. Continue picking a point at random and add it to the cluster whose mean it is nearest and updating the mean of the receiving cluster to take account of the new point until the entire data set is grouped into K clusters. The K principal points can then be taken to be the means of the K clusters.

1.5 Analytical Results

1.5.1 Self–Consistent Points

A set of points $y_1, y_2, \ldots, y_K$ is called self–consistent with respect to a p–dimensional random vector X if

$$E\{X|\delta_k(X, y_1, y_2, \ldots, y_K) = 1\} = y_k. \tag{1.16}$$

The sample version of self–consistent points is obtained by putting $X = x_U$ where U is distributed uniformly on $\{1, 2, \ldots, n\}$. A set of points is self–consistent if

$$y_k = \frac{\sum_i \delta_k(x_i, y_1, y_2, \ldots, y_K)x_i}{\sum_i \delta_k(x_i, y_1, y_2, \ldots, y_K)}, \quad \text{for all} \quad k = 1, 2, \ldots, K. \tag{1.17}$$

Some of the results for self–consistent sets of points are:

1. (Flury, 1993). The set of principal points is self–consistent.

2. The K–means clustering algorithm always converges to a set of self–consistent points.

3. (Tarpey, 1992). Let X denote a p–dimensional random vector with zero mean. Suppose $y_1, y_2, \ldots, y_K$ are self–consistent points of X and these points span a subspace of dimension $q < p$. Let $\alpha_1, \alpha_2, \ldots, \alpha_q \in \Re^p$ denote an orthonormal basis of this subspace, and set $B = (\alpha_1, \alpha_2, \ldots, \alpha_q)$. Then the random vector $B'X$ has a set of K self–consistent points $B'y_1, B'y_2, \ldots, B'y_K$.

4. (Tarpey et al, 1995). Let X_1 denote a p–dimensional random vector, and let $X_2 = \delta + \rho H X_1$ for some $\delta \in \Re^p, \rho \in \Re$ and for some $p \times p$ orthogonal matrix H. If $\{y_1, y_2, \ldots, y_K\}$ is a set of K self–consistent points of X_1, then $\delta + \rho H y_k$, $k = 1, 2, \ldots, K$ form a set of self–consistent points of X_2.

5. (Tarpey et al, 1995). Let X be p–dimensional vector with $y_1, y_2, \ldots, y_K$ as self–consistent points. Now if K self–consistent points of X span a subspace of dimension $q < p$, then this subspace is also spanned by the first q principal components.

1.5.2 Principal Points

1. (Tarpey et al, 1995). Let X_1 denote a p–dimensional random vector, and let $X_2 = \delta + \rho H X_1$ for some $\delta \in \Re^p, \rho \in \Re$ and for some $p \times p$ orthogonal matrix H. If $\{\xi_1, \xi_2, \ldots, \xi_K\}$ are principal points of X_1, then

$\delta + \rho H\xi_k$, $k = 1, 2, \ldots, K$, are principal points of X_2. Denote $\delta + \rho H\xi_k = \tilde{\xi}_k$, $k = 1, 2, \ldots, K$. Then $L_{F_2}\left\{\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_K\right\} = \rho^2 L_{F_1}\{\xi_1, \xi_2, \ldots, \xi_K\}$ where L_{F_1}, L_{F_2} are the loss functions defining principal points for the random variables X_1, X_2 with distribution functions F_1, F_2 respectively.

2. (Flury, 1990). Let X be a continuous univariate random variable with mean, $\mu = E(X)$, symmetrical density function $f(x)$, and with second finite moments. Then the $K = 2$ principal points are

$$\mu \pm E(|X - \mu|)$$

if and only if $f(\mu)E(|X - \mu|) < \frac{1}{2}$.

   (a) **Example.** Let X denote a univariate standard normal variable, and $f(x)$ its density. Then $f(\mu)E(|X-\mu|) = f(0)E(|X|) = \pi^{-1} \approx 0.318 < \frac{1}{2}$. Hence the two principal points of the standard normal are $\xi_1 = -(2/\pi)^{\frac{1}{2}} = -0.7977$ and $\xi_2 = (2/\pi)^{\frac{1}{2}} = 0.7977$. It follows that for $X \sim N(\mu, \sigma^2)$ and $K = 2$, the principal points are $\mu \pm \sigma(2/\pi)^{\frac{1}{2}}$ (Cox, 1957).

   (b) **Example.** For a uniform distribution on $[\mu - \eta, \mu + \eta]$ for $\eta > 0$, the two principal points are $\mu \pm \frac{1}{2}\eta$.

3. (Flury, 1990). Suppose a p–dimensional random vector X follows an elliptical distribution with mean vector μ and covariance matrix Σ. Then the two principal points of X have the form

$$\xi_k = \mu + \gamma_k \beta, \quad \text{for } k = 1, 2$$

where $\beta \in \Re^p$ is the normalized characteristic vector associated with the largest root of Σ, and γ_1, γ_2 are the two principal points of the univariate variable $\beta'(X - \mu)$.

4. (Flury, 1990). Suppose $X \sim N_p(\mu, \Sigma)$ and let ω_1 and β denote the largest characteristic root and vector, respectively of Σ. If ω_1 is simple, the two principal points of X are

$$\mu \pm (2\omega_1/\pi)^{\frac{1}{2}}\beta.$$

If ω_1 has multiplicity r $(1 \leq r \leq p)$, let $V^\star$ denote the sphere of radius $(2\omega_1/\pi)^{\frac{1}{2}}$, centered at μ, in the latent space of ω_1. Then any two points on $V^\star$ symmetric to μ are two principal points of X.

5. (Flury, 1990). Let $\mu \in \Re^p$, and Σ be a positive definite symmetric $p \times p$ matrix. Suppose a random vector X has a uniform distribution inside the set

$$C = \left\{x \in \Re^p : (x-\mu)^T\Sigma^{-1}(x-\mu) \leq 1\right\}.$$

Let ω_1 and β denote the largest characteristic root and associated characteristic vector, respectively of Σ, with $\beta'\beta = 1$. Suppose that ω_1 is simple. Then the two principal points of X are

$$\mu \pm \frac{2\omega_1^{\frac{1}{2}}\Gamma(\frac{1}{2}p+1)}{(p+1)\Gamma(\frac{1}{2}p+\frac{1}{2})\Gamma(\frac{1}{2})}\beta$$

provided that the local minimum taken at these points is the global minimum.

1.5.3 Principal Components

The reason for this section is to give a connection between principal points and principal components.

Let X be a p-dimensional random vector. Then the principal components are particular linear combinations of X. Principal component analysis depends

mainly on the covariance matrix Σ of X. The first principal component is the linear combination with maximum variability and last principal component is the linear combination with minimum variability. When the last $K+1, K+2, \ldots, n$ principal components have variances which can be considered as not significant, then principal components can be seen as a way of dimension or data reduction.

Now the result which gives the connection between principal points and principal components is given as:

1. (Tarpey et al, 1995). Suppose X is p–dimensional elliptical with $E(X) = 0$ and $Cov(X) = \Sigma$. If a set of K principal points of X spans a subspace V^* of dimension q, then Σ has a set of eigenvectors $\gamma_1, \gamma_2, \ldots, \gamma_p$ with associated ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ such that V^* is spanned by $\gamma_1, \gamma_2, \ldots, \gamma_q$.

If the eigenvalues of Σ are distinct then this result implies that if the K principal points of X span a subspace of dimension q, then this subspace is identical with the subspace spanned by the q principal components (q eigenvectors) associated with the q largest eigenvalues of Σ.

Chapter 2

Smoothing

2.1 Definitions and Rationale

Let X be a p–dimensional random vector having distribution function F. Let $\hat{F}$ be an estimate of F. This chapter investigates the use of smoothed estimates $\hat{F}$ to estimate principal points .

Let $\left\{\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K\right\}$ be an estimator of a set of K principal points. In evaluating the estimators, we will adopt the loss function $L_F\left(\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K\right)$. We get the risk function by taking the expectation of the loss using the true distribution function F :

$$R(\hat{F}, F) = E\left\{L_F\left(\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K\right)\right\} \tag{2.1}$$

$$= E\left\{\min_{1 \le k \le K} \left\|X - \hat{\xi}_k\right\|^2\right\}, \tag{2.2}$$

where X has distribution F and is independent of $\left\{\hat{\xi}_k\right\}$. We will evaluate the performance using the risk criterion.

This criterion seems to be more reasonable and tractable than a criterion based on the distance between $\left\{\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K\right\}$ and the actual principal points

$\{\xi_1, \xi_2, \ldots, \xi_K\}$ of X after a suitable ordering of the points. We must note that the function L_F may not have a unique minimum.

Define a smoothed empirical density function $\hat{f}_\tau$ to be the kernel–smoothed density estimate using the radial symmetric Gaussian kernel :

$$\hat{f}_\tau(x) = \frac{1}{n}\sum_{i=1}^{n}\left(2\pi\tau^2\right)^{-p/2}\exp\left\{-\frac{1}{2\tau^2}\|x - x_i\|^2\right\}. \tag{2.3}$$

Let $\hat{F}_\tau$ be the corresponding estimate of the distribution function, which can be written in short form notation as a mixture of n normals :

$$\hat{F}_\tau = \frac{1}{n}\sum_{i=1}^{n} N_p\left(x_i, \tau^2 I_p\right). \tag{2.4}$$

Note that $\hat{F}_0$ is the empirical distribution function, which puts mass $1/n$ at x_i.

When applying vector quantization to minimize $L_{\hat{F}_\tau}$, the optimality criterion (2.5), we will generate random vectors from $\hat{F}_\tau$. The procedure for generation of the random vectors is outlined in algorithm 2.1 as follows :

Algorithm 2.1

1. Generate a random variable W distributed uniformly on $\{1, 2, \ldots, n\}$.
2. Generate a p–dimensional vector Z of independent standard normal random variables; i.e. $Z \sim N_p(0, I_p)$, W and Z are independent.
3. Put $X = x_W + \tau Z$.

Now consider minimizing

$$\begin{aligned} L_{\hat{F}_\tau}(y_1, y_2, \ldots, y_K) &= E_{\hat{F}_\tau}\left\{\min_{1\le k\le K}\|X - y_k\|^2\right\} \\ &= E\left\{\min_k \|x_W + \tau Z - y_k\|^2\right\} \\ &= E\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\min_k \|x_i + \tau Z - y_k\|^2\right\}\right] \\ &= E\left\{L_{\hat{F}_0}(y_1 - \tau Z, \ldots, y_K - \tau Z) \mid \hat{F}_0\right\}. \end{aligned} \tag{2.5}$$

Note that the estimators $\left\{\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K\right\}$ minimize equation (2.5). Then $L_{\hat{F}_\tau}$ averages $L_{\hat{F}_0}$ as the configuration of vectors $y_1, y_2, \ldots, y_K$ is shifted by the vector τZ. Hence $L_{\hat{F}_\tau}$ is a smoother function of $y_1, y_2, \ldots, y_K$.

2.2 Optimal Smoothing for the Spherical Normal

In this section, we investigate the optimal values for the smoothing parameter τ when $F = N_p(0, I_p)$, for various choices of n, p, and K. We consider the distribution $F = N_p(0, I_p)$ because we know the optimal smoothing parameter (window width) τ_{DE} for the density estimation criterion (Silverman, 1986, pages 86–87):

$$\tau_{DE} = \left\{\frac{(2p+1)n}{4}\right\}^{-1/(p+4)}. \tag{2.6}$$

For this particular distribution, how does the risk equation (2.2) relate to (τ, n, p, K)? The value of τ for which the risk function is minimized will be called the optimal smoothing value and denoted τ_{PP}. How is τ_{PP} related to (n, p, K)? How much improvement in the risk is attained using $\tau = \tau_{PP}$ compared with $\tau = 0$, and how is this related to (n, p, K)? Answers to these questions for a specific distribution may provide insight concerning estimation of τ_{PP} from a data set with unknown distribution.

Now for different combinations and values of n, p, K we estimate the principal points and evaluate the risk function. The procedure is outlined as follows:
We start by specifying some values of τ with the first value $\tau_0 = 0$ and the upper bound $\tau_7 = 2\tau_{DE}$, where τ_{DE} is given as equation (2.6). Compute values of τ

from

$$\tau_i = \frac{2i}{7}\tau_{DE}\,, \quad \text{for } i = 0, 1, 2, \ldots, 7. \tag{2.7}$$

Estimation of the principal points and the loss is given in a "Do loop" below.

Do $i = 0$ to 7

Do $j = 1$ to 10

1. Generate a random sample $\{x_1, x_2, \ldots, x_n\}$ from $F = N_P(0, I_p)$. The samples are independent for different (i, j).

2. Apply vector quantization to estimate the principal points using $\tau = \tau_i$ denoted as $\left\{\hat{\xi}_1^\tau, \hat{\xi}_2^\tau, \ldots, \hat{\xi}_K^\tau\right\}$. i.e.

$$\left\{\hat{\xi}_1^\tau, \hat{\xi}_2^\tau, \ldots, \hat{\xi}_K^\tau\right\} = \underset{\xi_1^\tau, \xi_2^\tau, \ldots, \xi_K^\tau}{\operatorname{argmin}} E\left\{\min_{1 \le k \le K} \|xW + \tau Z - \xi_k^\tau\|^2\right\}, \tag{2.8}$$

where W and Z are defined as in algorithm 2.1.

3. Generate an independent sample $\{y_1, y_2, \ldots, y_{100,000}\}$ from $N_p(0, I_p)$ and estimate the loss L_F by

$$L_{ij} = \frac{1}{100,000} \sum_{l=1}^{100,000} \min_{1 \le k \le K} \left\|y_l - \hat{\xi}_k^\tau\right\|^2 . \tag{2.9}$$

End do.

End do.

In the "Do loop" above, we have j going from 1 to 10. The reason for this repetition is to get a more accurate risk estimate for each combination of n, p, and K.

Now for $\tau = \tau_i$, we evaluated the risk ten times hence we get eighty different values of the risk $L_F\left\{\hat{\xi}_1^\tau, \hat{\xi}_2^\tau, \ldots, \hat{\xi}_K^\tau\right\}$ each corresponding to some $\tau = \tau_i$. Hence the average, $\bar{\tau}$ of the τ's as

$$\bar{\tau} = \frac{1}{8}\sum_{i=0}^{7} \frac{2i}{7}\tau_{DE} = \tau_{DE}.$$

We then approximated the relationship between τ and L_F with a quadratic model. Now for fixed p, n, and K, model L_{ij} as

$$L_{ij} = \beta_0 + \beta_1(\tau_i - \bar{\tau}) + \beta_2(\tau_i - \bar{\tau})^2 + \varepsilon_{ij}. \tag{2.10}$$

Fitting a model of this form makes the least squares estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ uncorrelated.

Then from basic calculus, we differentiate equation (2.10) with respect to τ and set to zero. Hence we obtain the estimate of the optimal smoothing parameter τ_{PP} given as

$$\tau_{PP} = \hat{\tau} = \bar{\tau} - \hat{\beta}_1 / \left(2\hat{\beta}_2\right). \tag{2.11}$$

We computed two confidence intervals for each optimal smoothing value τ_{PP}. The first interval is an exact confidence interval and the second confidence interval is an approximation. In most cases, the two intervals were similar.

Now to compute the first interval for τ_{PP}, we consider the problem of testing the hypothesis

$$H_o : \frac{\beta_1}{2\beta_2} = C \qquad \text{or} \qquad H_o : \beta_1 - 2C\beta_2 = 0; \tag{2.12}$$

where C is some constant. Then the test statistics will be

$$t_c = \frac{\hat{\beta}_1 - 2C\hat{\beta}_2}{\left(S^2_{\hat{\beta}_1} + 4C^2 S^2_{\hat{\beta}_2}\right)^{1/2}} \tag{2.13}$$

where $S^2_{\hat{\beta}_1}$ and $S^2_{\hat{\beta}_2}$ are the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively. Then accept H_o if $\mid t_c \mid < t_{\alpha/2,\nu}$, $\nu = 80 - 2 - 1 = 77$ is the number of degrees of freedom. Let $t_{\alpha/2,\nu} = t_*$. We then proceed to solve for C from $\mid t_c \mid^2 < t_*^2$ as

$$\frac{\left(\hat{\beta}_1 - 2C\hat{\beta}_2\right)^2}{\left(S^2_{\hat{\beta}_1} + 4C^2 S^2_{\hat{\beta}_2}\right)} \leq t_*^2$$

$$\hat{\beta}_1^2 + 4C^2\hat{\beta}_2^2 - 4C\hat{\beta}_1\hat{\beta}_2 - t_*^2 S_{\hat{\beta}_1}^2 - 4C^2 t_*^2 S_{\hat{\beta}_2}^2 \leq 0$$

$$C^2\left(4\hat{\beta}_2^2 - 4t_*^2 S_{\hat{\beta}_2}^2\right) - C\left(4\hat{\beta}_1\hat{\beta}_2\right) + \left(\hat{\beta}_2^2 - t_*^2 S_{\hat{\beta}_1}^2\right) \leq 0.$$

This usually produces an interval for C with end points

$$\frac{4\hat{\beta}_1\hat{\beta}_2 \pm \sqrt{16\hat{\beta}_1^2\hat{\beta}_2^2 - 4\left(\hat{\beta}_1^2 - t_*^2 S_{\hat{\beta}_1}^2\right)\left(4\hat{\beta}_2^2 - 4t_*^2 S_{\hat{\beta}_2}^2\right)}}{2\left(4\hat{\beta}_2^2 - 4t_*^2 S_{\hat{\beta}_2}^2\right)}. \tag{2.14}$$

Let these two values be C_1 and C_2, where $C_1 < C_2$. We thus have a $(1-\alpha)\%$ confidence interval

$$C_1 \leq \beta_1/(2\beta_2) \leq C_2. \tag{2.15}$$

Now $\tau_{PP} = \bar{\tau} - \hat{\beta}_1/\left(2\hat{\beta}_2\right)$ so a confidence interval for τ_{PP}

$$\bar{\tau} - C_2 \leq \tau_{PP} \leq \bar{\tau} - C_1. \tag{2.16}$$

The second confidence interval is obtained by evaluating the approximate variance of $\hat{\beta}_1/\left(2\hat{\beta}_2\right)$ and then obtain the confidence interval for $\hat{\beta}_1/\left(2\hat{\beta}_2\right)$. After computing this interval, we then obtain the confidence interval for τ_{PP}. This is as follows. First let $f\left(\hat{\beta}_1, \hat{\beta}_2\right) = \hat{\beta}_1/\hat{\beta}_2$. Then the variance of $\hat{\beta}_1/\left(2\hat{\beta}_2\right) = \frac{1}{4}Var\left(\hat{\beta}_1/\hat{\beta}_2\right) = \frac{1}{4}Var\left\{f\left(\hat{\beta}_1, \hat{\beta}_2\right)\right\}$. Using linearization method we have

$$\begin{aligned}
f\left(\hat{\beta}_1, \hat{\beta}_2\right) &\approx f(\beta_1, \beta_2) + \left(\hat{\beta}_1 - \beta_1\right)\frac{\partial f}{\partial \beta_1} + \left(\hat{\beta}_2 - \beta_2\right)\frac{\partial f}{\partial \beta_2} \\
Var\left\{f\left(\hat{\beta}_1, \hat{\beta}_2\right)\right\} &\approx Var\left(\hat{\beta}_1\right)\left(\frac{\partial f}{\partial \beta_1}\right)^2 + Var\left(\hat{\beta}_2\right)\left(\frac{\partial f}{\partial \beta_2}\right)^2 \\
&\approx Var\left(\hat{\beta}_1\right)\left(\frac{\partial f}{\partial \hat{\beta}_1}\right)^2 + Var\left(\hat{\beta}_2\right)\left(\frac{\partial f}{\partial \hat{\beta}_2}\right)^2 \\
&= Var\left(\hat{\beta}_1\right)\left(\frac{1}{\hat{\beta}_2}\right)^2 + Var\left(\hat{\beta}_2\right)\left(\frac{\hat{\beta}_1}{\hat{\beta}_2^2}\right)^2 \\
&= A, \text{ say.}
\end{aligned}$$

So knowing that $\hat{\tau} = \bar{\tau} - \hat{\beta}_1/\left(2\hat{\beta}_2\right) = \bar{\tau} - \frac{1}{2}f\left(\hat{\beta}_1, \hat{\beta}_2\right)$ gives us $Var(\hat{\tau}) \approx A/4$. Hence if we let the standard error of $\hat{\tau} = S_\tau = \sqrt{A}/2$, then an approximate confidence interval for τ_{PP}

$$\tau_{PP} \in \hat{\tau} \pm 2S_\tau = \hat{\tau} \pm \sqrt{A}, \tag{2.17}$$

Now the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ used in the two cases (computing the intervals) are computed in the following manner. We first rewrite equation (2.10) in the form

$$L_F = \beta X' + \varepsilon$$

where $\beta = (\beta_0, \beta_1, \beta_2)$ and $X = [1, (\tau - \bar{\tau}), (\tau - \bar{\tau})^2]$. Then from regression analysis, we have $Var\left(\hat{\beta}\right) = \sigma^2 (X'X)^{-1}$ which is a 3×3 matrix given as

$$Var\left(\hat{\beta}\right) = \begin{bmatrix} Var\left(\hat{\beta}_0\right) & Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) & Cov\left(\hat{\beta}_0, \hat{\beta}_2\right) \\ Cov\left(\hat{\beta}_1, \hat{\beta}_0\right) & Var\left(\hat{\beta}_1\right) & Cov\left(\hat{\beta}_1, \hat{\beta}_2\right) \\ Cov\left(\hat{\beta}_2, \hat{\beta}_0\right) & Cov\left(\hat{\beta}_2, \hat{\beta}_1\right) & Var\left(\hat{\beta}_2\right) \end{bmatrix}.$$

Note that $Cov\left(\hat{\beta}_1, \hat{\beta}_2\right) = 0 = Cov\left(\hat{\beta}_2, \hat{\beta}_1\right)$ as $\hat{\beta}_1$ and $\hat{\beta}_2$ are defined to be uncorrelated. The estimate of σ^2 can be the maximum likelihood estimate which is the sum of squares of residuals divided by the sample size.

The design considered for the evaluation of the loss function L_F is given in Table 2.1. This design gave $5 \times 5 \times 5$ combinations of p, K, n. So we obtain 125 different values of the optimal smoothing parameter τ_{PP}.

Plots of τ_{PP} (not shown) indicate that τ_{PP} is an increasing function of K for fixed (p, n), a decreasing function of n for fixed (p, K), and decreasing function of p for fixed (n, K). All the values τ_{PP} were between 0 to 1.

After some exploratory analysis, we notice that τ_{PP}/τ_{DE} has a curvilinear relationship in (n, p, K) with an interaction between p and K. Plots of τ_{PP}/τ_{DE}

indicate that τ_{PP}/τ_{DE} is an increasing concave function of K and a decreasing convex function of p shown as Figure 2.1(b) and Figure 2.2 respectively. We model τ_{PP}/τ_{DE} as the sum of a linear function of $1/\sqrt{n}$ and a quadratic function of (p, K). The model equation obtained is

$$\tau_{PP}/\tau_{DE} = 0.475 + \frac{1.529}{\sqrt{n}} + 0.070K - 0.117p - 0.001K^2 + 0.009p^2 - 0.003pK. \tag{2.18}$$

The fitted model gave $R^2 = 92\%$. The standard error of regression coefficients in equation (2.18) is given in Table 2.2. Residual analysis of the fit suggest that the residual variance is constant as n and p vary but is a decreasing function of K. The graph of residuals versus K is shown in Figure 2.1(a). A quantile plot of the residuals of the fit gives a reasonably straight line.

We now examine the reduction in the risk due to smoothing. Let the risk at $\tau = 0$ be $R_0 = R(\hat{F}_0, F)$ and the risk at $\tau = \tau_{PP}$ be $R_{opt} = \min_\tau R(\hat{F}_\tau, F)$. We consider the estimate of the risk ratio R_{opt}/R_0. We note that R_0 and R_{opt} are estimates and are obtained by evaluating the quadratic (equation 2.10) at $\tau = 0$ and $\tau = \tau_{PP}$ respectively. The risk estimates obtained by computing the average of the loss L_F (equation 2.9) at $\tau = 0$ and $\tau = \tau_{PP}$ after estimation of the principal points at these points indicate that there is very little difference between the two risk estimates and there is no negative bias in the two risk estimates. This ratio ranged from 0.10 to 1.00, but most of the values were close to 1.00. This indicates that smoothing provides little improvement for most values of (n, p, K). Substantial improvement was evident for K large, n small, and p small. A plot of R_{opt}/R_0 versus a linear combination of n, p, and K is given in Figure 2.3 from which one can see that most of R_{opt}/R_0 values are close

to one. The linear combination of n, p, and K used is

$$\text{lincomb}(n, p, K) = 2.46n - 31.80K + 177.00p. \tag{2.19}$$

This linear combination was determined using a flexible exponential regression technique called normalized exponential smoothing, (Hooper, 1996). Hence as K increases with small n and p, substantial improvement with smoothing can be seen as displayed in Figure 2.3.

Table 2.1: Table for Design of Model.

p	1	2	3	5	10
K	2	3	5	10	20
n	25	50	100	200	500

Table 2.2: Table for the results for the model approximation.

Dependent variable		τ_{PP}/τ_{DE}		
R Squared		91.9 %		
s = 0.0747	125 - 7 = 118 degrees of freedom (df)			
Source	Sum of Squares	df	Mean Square	F-ratio
Regression	7.4727	6	1.2454	223
Residual	0.6589	118	0.0056	
Variable	Coefficient	S.e of Coef	t-ratio	Prob
Constant	0.4754	0.0319	14.90	< 0.0001
$1/\sqrt{n}$	1.5286	0.1221	12.50	< 0.0001
K	0.0697	0.0052	13.50	< 0.0001
p	-0.1172	0.0103	-11.30	< 0.0001
K^2	-0.0015	0.0002	-6.77	< 0.0001
p^2	0.0089	0.0009	10.30	< 0.0001
Kp	-0.0031	0.0003	-9.66	< 0.0001

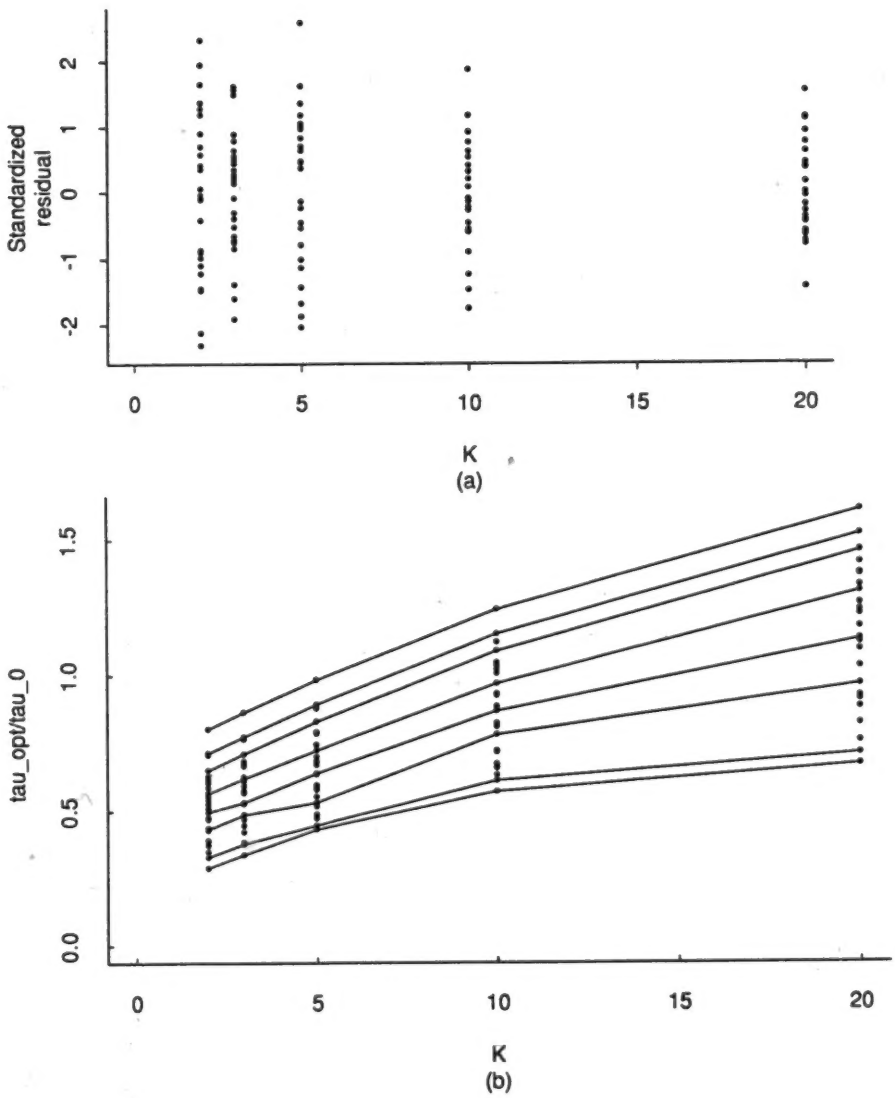


Figure 2.1: (a): Plot of residuals versus number of principal points, K and (b): Plot of τ_{PP}/τ_{DE} versus the number of principal points, K.

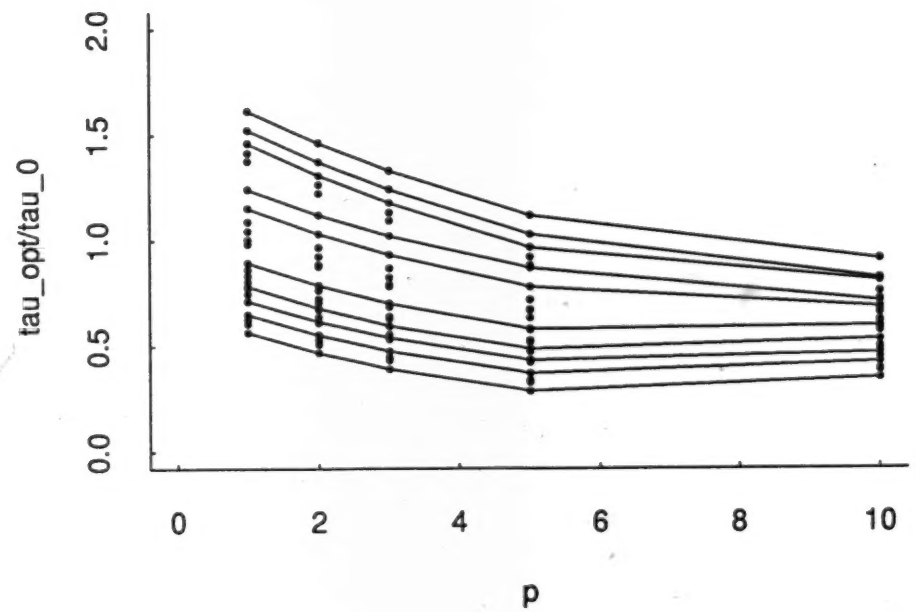


Figure 2.2: Plot of τ_{PP}/τ_{DE} versus the dimension, p.

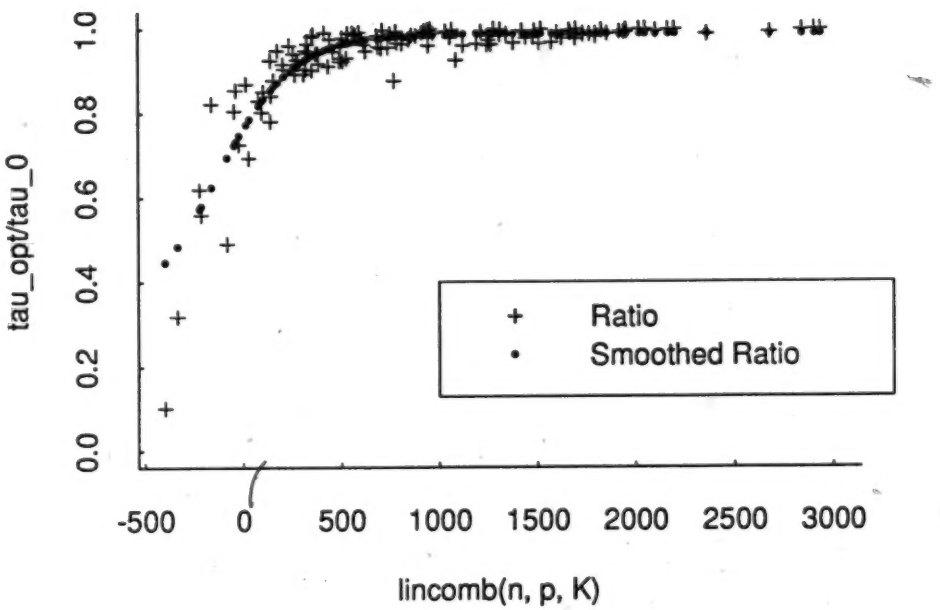


Figure 2.3: Plot of R_{opt}/R_0 versus linear combination of n, p, and K as given in (2.19).

Chapter 3

Estimation of Principal Points

3.1 Introduction

Applying the smoothing technique introduced in the previous chapter, we now consider how to estimate principal points for a data set. Our main interest in this chapter is to use cross–validation to estimate the risk and then obtain an estimate of the optimal smoothing value using the procedure in section 2.2. We begin by introducing the concept of v–fold cross–validation and then proceed to use it in estimation of the loss.

3.2 v–Fold Cross–Validation

Cross–validation (Stone, 1974) is a method for estimating prediction error in regression and classification problems. It has also been used for model selection in nonparametric applications.

The procedure in v–fold cross–validation is described below. We first partition the data set randomly into v disjoint groups of sizes $n_1, n_2, \ldots, n_v$, where $[n/v] \leq n_i \leq [n/v] + 1$. Hence groups have roughly equal sizes. Denote the

groups by $A_1, A_2, \ldots, A_v$ with sizes $n_1, n_2, \ldots, n_v$ respectively . Also let $A^{(-i)}$ be the subset which contains all observation excluding observations from A_i. Then $A^{(-i)}$ contains $(n - n_i)$ observations. We refer to $A^{(-i)}$ as a training set and A_i as a test set or group. In prediction problems, a prediction rule is developed on the observations in $A^{(-i)}$ and then this rule is tested on the observations in the test set A_i. i.e. evaluate the risk. This procedure is repeated for each $i = 1, 2, \ldots, v$. Then we get v different estimates of the risk. Finally take the average of these v risks. This average is called the v–fold cross–validated risk estimate. When $v = n$, it is called ordinary cross–validation. Observations are assumed to be independent. The difficulty with ordinary cross–validation is that it can be computationally very expensive, because an estimate has to be developed n times. In v–fold cross–validation, $v < n$, we need to develop the estimate only v times.

The description of the term "ordinary cross–validation" from Stone (1974) is given as follows :

> "Suppose we set aside one individual case, optimize for what is left, then test on the set–aside case. Repeating this for every case squeezes the data almost dry. If we have to go through the full optimization calculation every time, the extra computation may be hard to face. Occasionally we can easily calculate either exactly or to an adequate approximation what the effect of dropping a specific and very small part of the data will be on the optimized result. This adjusted optimized result can then be compared with the values for the omitted individual. That is, we make one optimization for all the data, followed by one repetition per case of a much simpler calculation, a calculation of the effect of dropping each individual,

followed by one test of that individual. When practical, this approach is attractive."

Stone (1974) argued that the ordinary cross–validation is asymptotically optimal.

Cross–validation looks like the jack–knifing procedure since both employ the method of omission of one or more observations. But the main difference is cross–validation deals with the problem of prediction whilst jack–knifing procedure deals with variance estimation and bias reduction.

3.3 Optimal Smoothing Estimated from a Data Set

In this section, we discuss how we can estimate the optimal value of τ. Let $\{x_1, x_2, \ldots, x_n\} \in \Re^p$ be a data set. Partition $\{1, 2, \ldots, n\}$ into subsets $A_1, A_2, \ldots, A_v$ of size $n_1, n_2, \ldots, n_v$ respectively just as in section 3.2. Recall from section 3.2 that $A^{(-i)}$ is the training set and the corresponding A_i is the test set. First get an estimate of the risk and then the optimal smoothing value just as in section 2.2.

Now the estimation of the principal points and evaluation of the risk is as follows. Start by applying vector quantization to estimate an initial set of principal points using the entire (unsmoothed) data set. Let these initial K principal points estimates be $\hat{\xi}_1^*, \hat{\xi}_2^*, \ldots, \hat{\xi}_K^*$. With the initial principal point estimates, compute a scale parameter estimate (the total distance between the nearest neighbors among the initial principal points estimates) :

$$t_{nn} = \sum_{k=1}^{K} \min_{l \neq k} \left\| \hat{\xi}_k^* - \hat{\xi}_l^* \right\|. \tag{3.1}$$

Note that this estimate t_{nn} is computed only once for each (n, p, K). Empirical evidence suggests that t_{nn} can be taken or considered as the upper bound for the smoothing parameter τ. Now define the smoothing parameter as

$$\tau = c_\tau t_{nn}, \tag{3.2}$$

where c_τ is some chosen constant. Now apply vector quantization to compute the principal points using the smoothed data as

$$x_W + \tau Z$$

from the training set $A^{(-i)}$, where W is a random variable distributed uniformly on $A^{(-i)}$, Z is a p–dimensional vector of independent standard normal variables, i.e. $Z \sim N_p(0, I_p)$, Z and W are independent, and with τ as in equation (3.2). Let this set of K principal point estimates be denoted $\hat{\Xi}_\tau^{(-i)} = \left\{\hat{\xi}_{\tau,1}^{(-i)}, \hat{\xi}_{\tau,2}^{(-i)}, \ldots, \hat{\xi}_{\tau,K}^{(-i)}\right\}$. Also let $\hat{F}^{(i)}$ be the empirical distribution based on A_i (unsmoothed). We then evaluate the risk function using the unsmoothed test set A_i. This is given as

$$L_{\hat{F}^{(i)}}\left\{\hat{\Xi}_\tau^{(-i)}\right\} = \frac{1}{n_i} \sum_{j \in A_i} \min_k \left\| x_j - \hat{\xi}_{\tau,k}^{(-i)} \right\|^2. \tag{3.3}$$

Repeat the procedure of estimating the principal points from the training set $A^{(-i)}$ and evaluation of the risk for all values of $i = 1, 2, \ldots, v$. This gives v different values for the risk estimate. Then the cross-validated risk estimate is the average of the v risks :

$$R(\tau) = \frac{1}{v} \sum_{i=1}^{v} L_{\hat{F}^{(i)}}\left(\hat{\Xi}_\tau^{(-i)}\right). \tag{3.4}$$

Now to get the optimal value of the smoothing parameter τ, we have to compute the cross–validated risk estimate for various values of c_τ. The values of

c_τ can be chosen in the following way. Start with the value $c_\tau = 0$ and obtain the risk estimate. Increase c_τ by a small increment and obtain the risk estimate. For example, we use increments of 0.025. Empirical evidence suggests that 0.025 is a good choice. Continue to increase the value of c_τ by the same amount and obtain the corresponding risk estimate until the risk estimate fails to decrease. Suppose the smallest value of the risk occurred at the mth value of c_τ. Then it follows that the risk fails to decrease at the $(m+1)$th value of c_τ. Because of the randomness in the computation (vector quantization algorithm), it is likely that the mth value of the risk obtained will not be the smallest risk. But the minimum risk will be a value close to the mth risk estimate. Hence obtain three or four more risk estimate after that minimum risk. i.e. obtain the $(m+2)$, $(m+3)$, and $(m+4)$th risk estimates. In this case, there is a high chance that the minimum risk will be within the range of the risk estimate obtained at the first c_τ value (which is 0) and the $(m+4)$th value. This computation is done with the same subsets $A_1, A_2, \ldots, A_v$ throughout. In our computations, we observe that the minimum risk estimate usually occurs within the first five values of c_τ.

Hence we will now be able to get an estimate of the optimal value of τ after fitting a quadratic model just as in section 2.2. Proceed as follows. First approximate the relationship between $R(\tau)$ (given as in equation 3.4) and τ by fitting a quadratic model as in (2.10) with L_{ij} replaced by $R(\tau)$:

$$R(\tau) = \beta_0 + \beta_1(\tau - \bar{\tau}) + \beta_2(\tau - \bar{\tau})^2 + \varepsilon. \tag{3.5}$$

Then estimate the optimal smoothing value τ_{PP} by

$$\tau_{PP} = \hat{\tau} = \bar{\tau} - \hat{\beta}_1 / \left(2\hat{\beta}_2\right) \tag{3.6}$$

which is the same as equation (2.11).

Chapter 4

Numerical Examples

4.1 Representative Curves

Our first example is one which has been studied by Johns and Rice, 1992; Flury and Tarpey, 1993. The problem they all studied was how to select a set of representative curves from a collection of curves.

On Figure 4.1 are superimposed a collection of 100 similar curves generated by an application of a fixed procedure to each of 100 data sets simulated from a given model. From Figure 4.1, we can see that many lines overlap and it is very difficult to examine variation among the various curves. We utilize principal points estimated by application of vector quantization to select representative curves from the original set. In this example, each of the 100 curves on Figure 4.1 is a kernel density estimate (e.g., Silverman, 1986)

$$\hat{f}(x) = \frac{1}{50}\sum_{i=1}^{50}\frac{1}{\tau_{DE}}\phi\left\{\frac{x - X_i}{\tau_{DE}}\right\} \tag{4.1}$$

based on a different sample $X_1, X_2, \ldots, X_{50}$ from the standard normal distribution (ϕ is the standard normal density). The value of the smoothing parameter, τ_{DE} in equation (4.1) is obtained from equation (2.6) (Silverman, 1986) by

putting $p = 1$ and $n = 50$. We select for display a few curves representing the 100 on Figure 4.1. We investigate the effect of smoothing (choosing $\tau > 0$) on this representation.

Evaluate each of the 100 curves at some equally spaced points. We evaluated at 101 equispaced points. In this way, we obtain a data set of 100 vectors of dimension 101. Apply vector quantization to estimate the K principal points corresponding to the values of $\tau = 0$ and $\tau = \tau_{PP}$, plot the coordinates and interpolate to obtain K curves (Flury and Tarpey, 1993). Note that we first have to estimate the value of τ_{PP} from this data set.

A 10–fold cross–validation was used to estimate the risk. Listed in Table 4.1 are the ratios of the risk estimates for $\tau = 0$ and $\tau = \tau_{PP}$ with their corresponding values of τ_{PP}. Values of K used in the computations are $K = 2$, 3, 4, 5, 7, 8, and 10. Principal points for $K = 2$, 3, 4, and 5 shown as Figures 4.2(a), 4.2(b), 4.3(a), and Figure 4.3(b) respectively. The graph corresponding to the value of $\tau = 0$ is indicated in all cases as dash lines and those corresponding to the value $\tau = \tau_{PP}$ as solid lines. There appears to be little difference between the graphs corresponding to the two values of τ. They almost coincide, especially in the case $K = 2$, but as K increases some difference can be seen.

Table 4.1 lists values of τ_{PP} for various K. As anticipated from results in Chapter 2, τ_{PP} tends to increase with K. Table 4.1 also provides two estimates of the ratio of the risk for $\tau = 0$ and $\tau = \tau_{PP}$. The estimates R_0 and R_{opt} were obtained by substituting $\tau = 0$ and $\tau = \tau_{PP}$ into the quadratic equation (3.5), where the coefficients were estimated by the technique of section 3.3. The estimates R_0^* and R_{opt}^* were obtained by 30–fold cross–validation for $\tau = 0$ and $\tau = \tau_{PP}$.

The estimates R_{opt}/R_0 and R^*_{opt}/R^*_0 yield substantially different results. The first ratio indicates that smoothing produces a substantial reduction in the risk when τ_{PP} is large. The second ratio indicates that the improvement is slight. I would argue that the second is more accurate. This is based in part on the plots in Figures 4.2 and 4.3, and on the following consideration. In approximating the risk function $R(\tau)$ by a quadratic, it is possible that $R(0)$ is overestimated and $R(\tau_{PP})$ is underestimated, producing negative bias in R_{opt}/R_0 .

The curves corresponding to $\tau = 0$ seem to be smoother than those for $\tau = \tau_{PP}$. This is likely due to the fact that when $\tau = 0$, the curves are just weighted averages of the original curves but when $\tau = \tau_{PP}$, small perturbations are introduced at equispaced points along the curve. This can be viewed as an argument against smoothing in this context.

4.2 Ozone Data

Our second example concerns recordings of daily maximum one–hour–average ozone levels in Upland, California, for the year 1976. Complete data set is given for only $n = 47$ weeks of that year. The remaining weeks have missing data so we use only the complete data weeks. The raw data set, plotted by linearly interpolating the $p = 7$ daily values for each week are given in Figure 4.4. High values of the ozone level were recorded during the summer (May—October) and low values in the winter (November—April). Some weeks in the summer are characterized with either decreasing ozone levels or increasing ozone levels or both.

Some of the authors who have studied this data set are Johns and Rice, (1992) and Flury and Tarpey, (1993). Johns and Rice (1992) selected representative

set of curves from a collection of curves using a principal component analysis to identify important modes of variation among the curves and principal component scores to identify particular curves for representation. To select say the curve with the rth greatest variation in a particular principal component axis, select the curve corresponding to $C_{(r)}$ the rth order statistics of the principal component scores where the $100\alpha\%$ quantile is defined as $C_{([n\alpha]+1)}$, $[x]$ as the integer part of x.

The problem is to select for display a few representative curves from a collection of curves. We utilize principal points estimated by application of vector quantization to select representative curves from the original curves. A 10–fold cross validation is used to compute the risk estimates. The risk estimates for $\tau = 0$ and $\tau = \tau_{PP}$ is reported as ratio of the two risk estimates with their corresponding τ_{PP} in Table 4.2. This is done for $K = 2$, 3, 4, 5, 7, 8, and 10. We produce the graphs of principal points for $K = 2$, 3, 4, and 5 shown as Figures 4.5(a), 4.5(b), 4.6(a) and Figure 4.6(b) respectively. For $K = 2$, we have a point representing the summer and one representing the winter. Also when $K = 3$, we have the same explanation for $K = 2$ but in addition, a points representing between winter and summer. Finally for $K > 3$, we have points which are essentially linear changes in levels through the week: some weeks there is an increase, in others a decrease.

The risk ratio estimate R_{opt}/R_0 again appears to be negatively biased, as it suggests improvement up to 34% while R^*_{opt}/R^*_0 indicates only 18%.

4.3 Boston Housing Data

Our third example concerns observations from census tracts in the Boston area originally studied by Harrison and Rubinfeld, (1978). Some of the authors who

have also studied this data set are Belsley et al. (1980), Breiman and Friedman (1985), and Hooper (1996). A listing of the data set is provided by Belsley et al. (1980). Fourteen variables were observed to describe each tract and they are :

X_1 = median value of owner–occupied home.

X_2 = crime rate by town. High crime rate is the inner–city areas and low in the suburban areas.

X_3 = proportion of town's residential land zoned for lots greater than 25,000 square feet. The proportion is low in the inner–city areas and high in the suburban areas.

X_4 = proportion of nonretail business acres per town. It is high in the inner–city areas and low in the suburban areas.

X_5 = Charles River dummy = 1 if tract bounds the Charles River, 0 otherwise.

X_6 = nitrogen oxide concentration (parts per hundred million). The concentration is high in the inner–city areas and low in the suburban areas.

X_7 = average number of rooms in owner units.

X_8 = proportion of owner–occupied units built prior to 1940. The proportion is high in the inner–city areas and low in the suburban areas.

X_9 = weighted distance to five employment centers in the Boston region. The weighted distance is small for the inner–city areas and high for the suburban areas.

X_{10} = index of accessibility to radial highways. The index is high for the inner–city areas and low for the suburban areas.

X_{11} = full–value property tax rate (per $ 10,000). The property tax rate is high in the inner–city areas and low in the suburban areas.

X_{12} = pupil–teacher ratio by town school district. It is high in the inner–city

areas and low in the suburban areas.

X_{13} = black proportion of population.

X_{14} = proportion of population that is lower economic status. The proportion is high for in the inner-city areas and low in the suburban areas.

The sample size for this example is 506 and dimension is 14. Since the variables were measure on different scales, we first center and standardize each of the variables to have unit variance and mean zero before computing the principal points. So the principal points here can be seen as a representative points of the standardized data. A 10–fold cross validation is used to compute the risk estimates. Listed in Table 4.3 are ratios of the risk estimates for $\tau = 0$ and $\tau = \tau_{PP}$ and their corresponding τ_{PP}. This is done for $K = 2$, 3, 4, 5, 7, 8, and 10. We produce the graphs of principal points for $K = 2$, 3, 4, and 5 shown as Figures 4.7(a), 4.7(b), 4.8(b) and Figure 4.8(b) respectively. When $K = 2$, the principal points are roughly symmetric and one point represents the inner–city areas with high crime, more industry, poor air quality, more old buildings, high tax, and more poverty. The second point is just the opposite of the inner–city areas, and it represent the suburban areas. For $K = 3$, two points are similar to those $K = 2$. Interpretation of the third point depends on whether smoothing was used. The unsmoothed principal point identifies tracts bounded by the Charles River. This is not the case with the smoothed principal points.

The risk ratio estimate R_{opt}/R_0 in this example indicates improvement up to 30% but R^*_{opt}/R^*_0 indicates only 15%. Again R_{opt}/R_0 appears to be negatively biased.

4.4 Seals Data

Our last example is concerned with harp seals (Phoca groenlandica), and in particular the herds from Jan Mayen Island, Gulf of St. Lawrence, and Front. Front represents the herd which is geographically located to the east of Newfoundland. A seal is a fish–eating mammal with four flippers which is aquatic but comes on shore to breed. Seals possess varied repertoires of underwater vocalisations (calls).

The data set is in the public domain by courtesy of Prof. J. M. Terhune, Department of Biology, University of New Brunswick. One thousand calls from each of the three herds were recorded, and several features of each recording were noted. There are eight variables in total. Complete data are available for only seven variables. These seven variables are :

X_1 :— the duration of a single element of a harp seal underwater vocalisation, measured in milliseconds.

X_2:— the number of elements of the call. In harp seals all of the elements within a single call are similar and the spacing between them is constant.

X_3 :— the pitch at the start of the call or the highest pitch if the call has an extremely short duration (call shape 0 below). This variable is measured in Hertz (Hz).

X_4 :— the pitch at the end of the call or the lowest pitch if the call has an extremely short duration (call shape 0). This variable is measured in Hertz (Hz).

X_5 :— codes a series of waveform shapes (a plot of amplitude versus time) which lie more or less along a continuum. The waveform shapes are: frequency modulated sinusoidal = 9, slight FM and complex = 8, sinusoidal (pure tone) = 7, complex (irregular waveform) = 5, amplitude pulses = 4, burst pulses = 3, knock

(short burst pulse) = 2, and click (very short duration) = 1.

X_6 :— this codes a series of call shapes as they would appear in a sonogram spectral analysis (a plot of frequency versus time). The shapes lie along a continuum.

X_7 :— this is the herd from which the recordings were obtained. This variable is not used in the computation of the principal points.

The sample size is 3000 and dimension is 7. Since the variables were measured on different scales, we first center and standardize each of the variables to have unit variance and mean zero before computing the principal points. We computed τ_{PP} and the risk for $K = 2$, 3, 4, 5, 7, 8, and 10. A 10–fold cross validation is used to compute the risk estimates. The results are listed in Table 4.4. We produce the graphs of principal points for $K = 2$, 3, 4, and 5 shown as Figures 4.9(a), 4.9(b), 4.10(a) and Figure 4.10(b) respectively. For $K = 2$, principal points are roughly symmetric. One principal points represents a pitch at the start of a call. The second is just the opposite of the first point. For $K = 3$, two of the points are similar to the case when $K = 2$. The third principal point shows a call with high number of elements and the waveform shape is click, i.e. has a very short duration. $K = 4$ and $K = 5$ have three of the principal points similar to that of $K = 3$. The smallest risk estimate occurs at $K = 7$, so the data set contains 7 clusters.

The risk ratio estimate R_{opt}/R_0 in our final example indicate improvements up to 37% but R^*_{opt}/R^*_0 indicates only 16%. Again R_{opt}/R_0 shows a negative bias.

Now Figure 4.11 displays the principal points for $K = 1$ and $K = 2$ for each of the three herds. Principal points from Jan Mayen Island is represented as solid

lines, Gulf of St. Lawrence represented as dotted lines, and Front as broken lines. When $K = 1$, it can be seen that the calls from Gulf of St. Lawrence, Jan Mayen Island, and Front does not show much difference. A multivariate analysis to test that the principal points of the three region are the same gives a p-value of 0. Hence we reject that hypothesis. When $K = 2$, the principal points all have differences.

4.5 Conclusion

We have considered the estimation of principal points and the risk function using the smoothing technique. In this approach, we have studied the optimal smoothing value when estimating principal points. First, we studied in Chapter 2 the relationship between the optimal smoothing values τ_{PP} and τ_{DE} using the spherical normal distribution. Vector quantization was used in the estimation of the risk. It was noted that substantial improvement with smoothing over not smoothing was evident for large K and small (n, p). In Chapter 3, we suggested a procedure for the estimation of the risk function using v–fold cross–validation. Finally in Chapter 4, we give some numerical examples illustrating the procedure suggested in Chapter 3.

In all of our examples, we observed that the estimate R_{opt}/R_0 appears to be negatively biased. The risk estimates R_0^* and R_{opt}^* appears to be more reliable. The improvement due to smoothing indicated by R_{opt}^*/R_0^* in the examples is small, and may not warrant the substantial computational effort required to select τ_{PP}.

Table 4.1: Results for Representative Curves.

$n = 100$ and $p = 101$.

R_{opt} is the estimate of the risk function evaluated at $\tau = \tau_{PP}$.

R_0 is the estimate of the risk function evaluated at $\tau = 0$.

R^*_{opt} is the 30–fold cross–validated estimate of the risk function at $\tau = \tau_{PP}$.

R^*_0 is the 30–fold cross–validated estimate of the risk function at $\tau = 0$.

K	τ_{PP}	R_{opt}/R_0	R^*_{opt}/R^*_0
2	0.02340	0.99741	0.99599
3	0.05905	0.96746	0.97521
4	0.08318	0.82954	0.86729
5	0.09636	0.72057	0.89979
7	0.10415	0.58642	0.83953
8	0.10296	0.61460	0.86360
10	0.08093	0.74186	0.84555

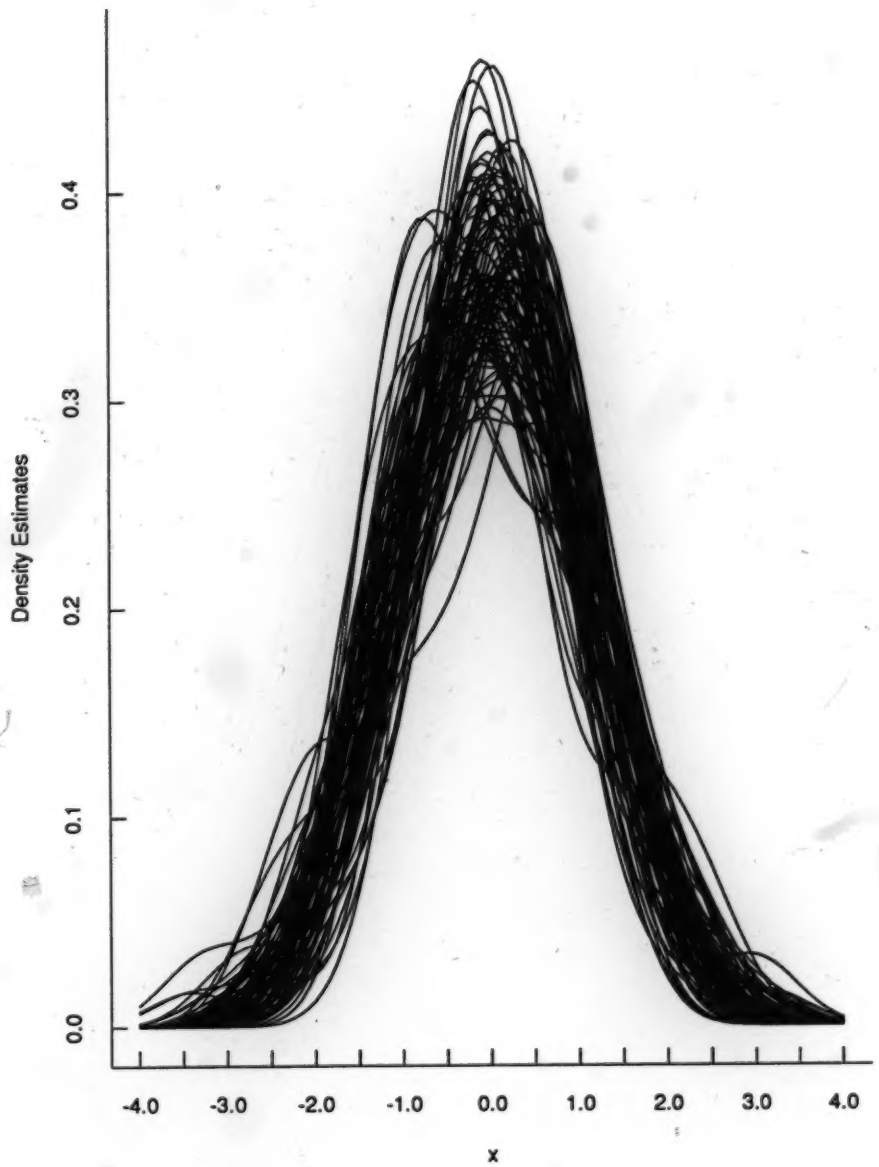


Figure 4.1: A collection of Density Estimates (Example 1). One hundred kernel density estimates were used based on independent samples of size fifty from the standard normal distribution overlaid on one another

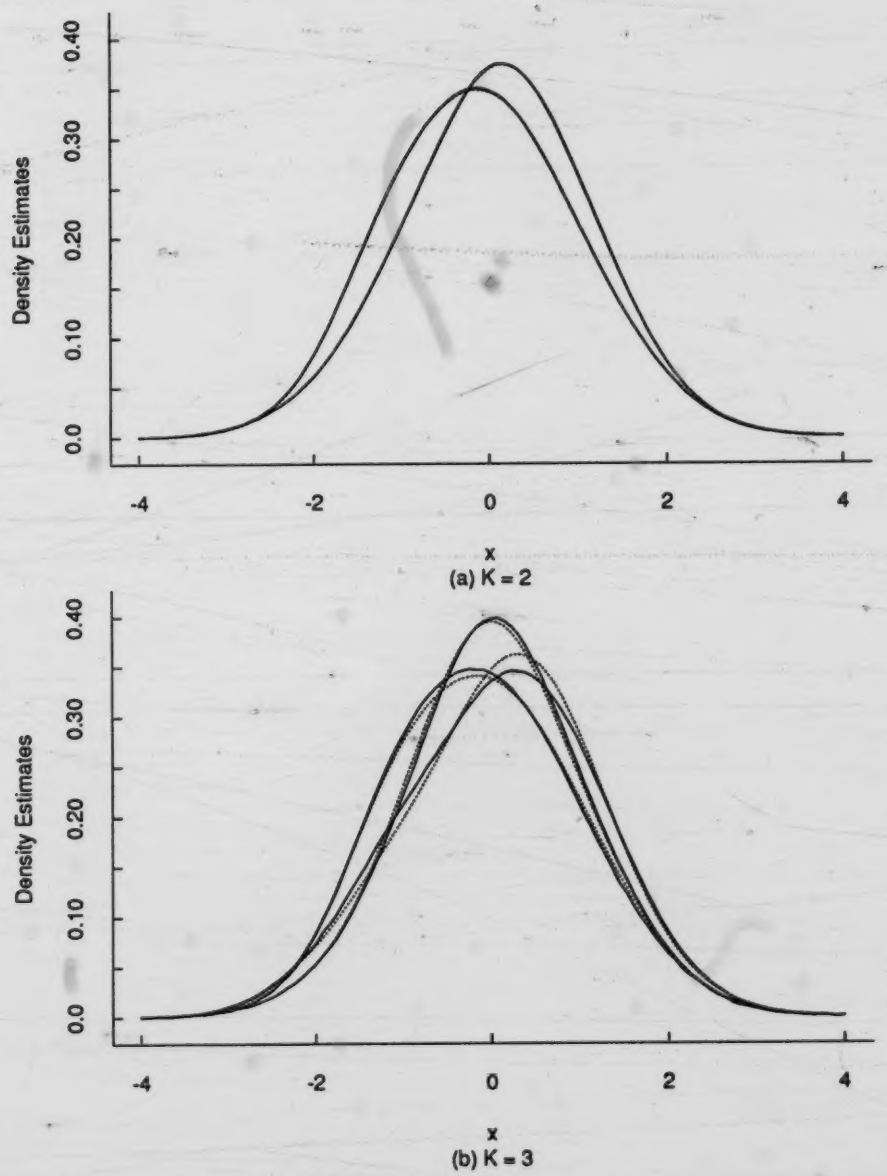


Figure 4.2: Representative Curves based on ($K = 2$ and 3) principal point estimates in Example 1 (Density Estimation). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines).

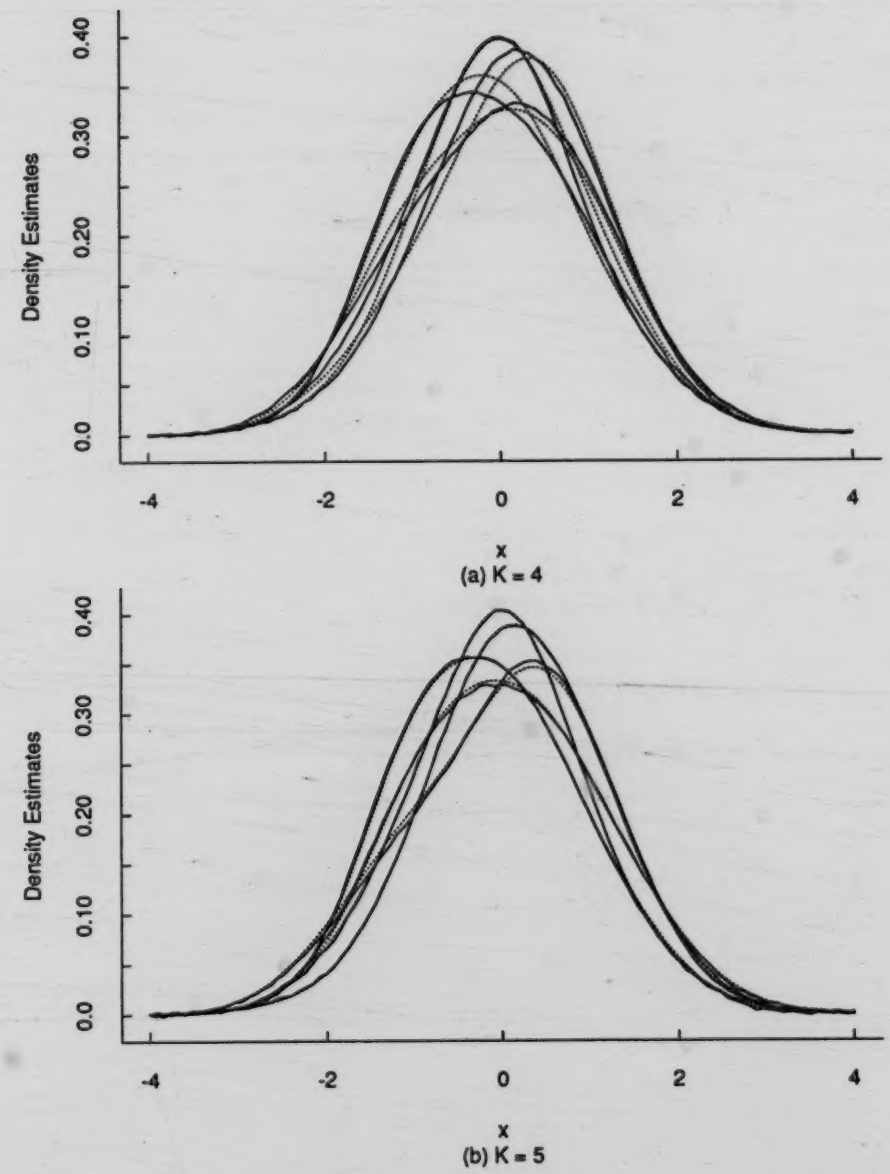


Figure 4.3: Representative Curves based on ($K = 4$ and 5) principal point estimates in Example 1 (Density Estimation). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines).

Table 4.2: Results for Ozone Data.

$n = 47$ and $p = 7$.

R_{opt} is the estimate of the risk function evaluated at $\tau = \tau_{PP}$.

R_0 is the estimate of the risk function evaluated at $\tau = 0$.

R^*_{opt} is the 30–fold cross–validated estimate of the risk function at $\tau = \tau_{PP}$.

R^*_0 is the 30–fold cross–validated estimate of the risk function at $\tau = 0$.

K	τ_{PP}	R_{opt}/R_0	R^*_{opt}/R^*_0
2	7.08604	0.95018	0.98624
3	3.54884	0.97733	0.97424
4	5.42396	0.86807	0.87099
5	5.74509	0.82478	0.90576
7	6.66955	0.72737	0.92901
8	6.70409	0.73252	0.83769
10	7.83980	0.65162	0.81202

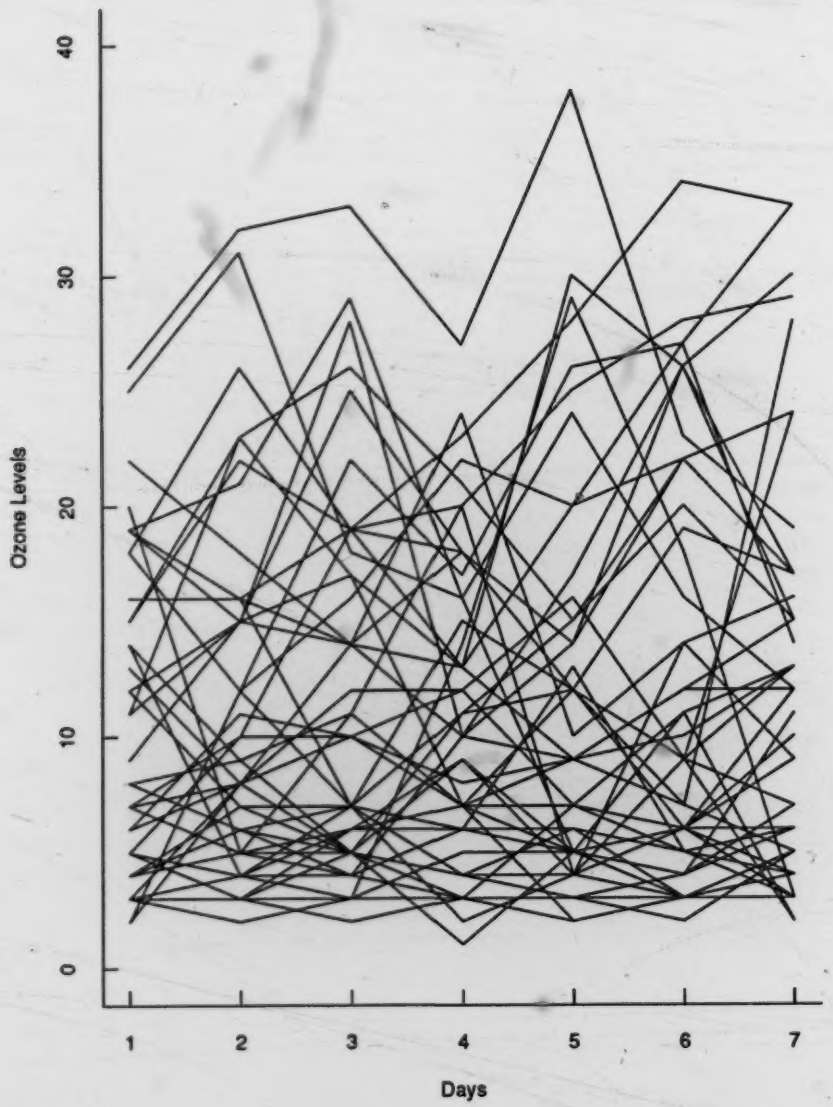


Figure 4.4: The Ozone Data (Example 2). Plot of ozone data by linearly interpolating values for each week. On the x–axis, 1 = Monday,...,7 = Sunday and ozone levels on the y-axis.

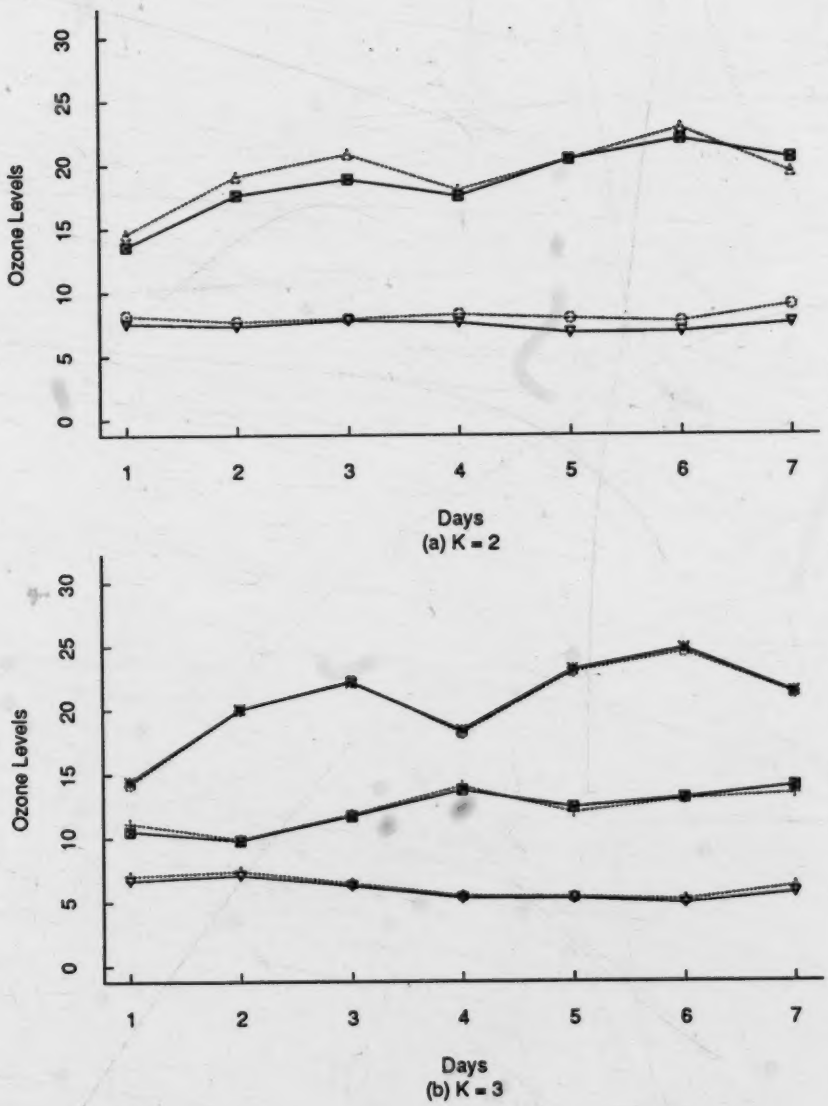


Figure 4.5: Representative Curves based on ($K = 2$ and 3) principal point estimates in Example 2 (Ozone data). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines). On the x-axis, 1 = Monday,...,7 = Sunday and ozone levels on the y-axis.

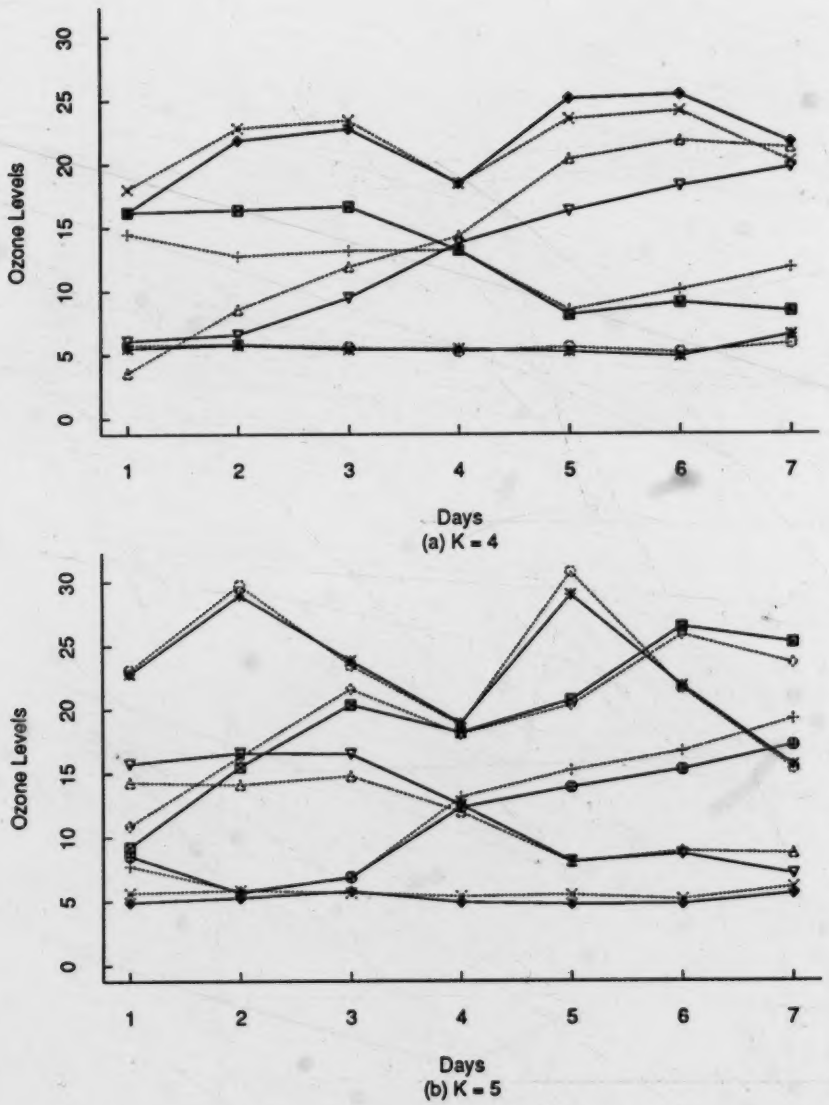


Figure 4.6: Representative Curves based on (K = 4 and 5) principal point estimates in Example 2 (Ozone data). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines). On the x-axis, 1 = Monday,...,7 = Sunday and ozone levels on the y-axis.

Table 4.3: Results for Boston Housing Data.

$n = 506$ and $p = 14$.

R_{opt} is the estimate of the risk function evaluated at $\tau = \tau_{PP}$.

R_0 is the estimate of the risk function evaluated at $\tau = 0$.

R^*_{opt} is the 30–fold cross–validated estimate of the risk function at $\tau = \tau_{PP}$.

R^*_0 is the 30–fold cross–validated estimate of the risk function at $\tau = 0$.

K	τ_{PP}	R_{opt}/R_0	R^*_{opt}/R^*_0
2	0.66550	0.98783	0.99784
3	0.53818	0.97767	0.98136
4	0.81781	0.90673	0.97745
5	0.93491	0.82602	0.95276
7	1.16061	0.71503	0.92271
8	1.15363	0.70384	0.90703
10	1.07624	0.7451578	0.84748

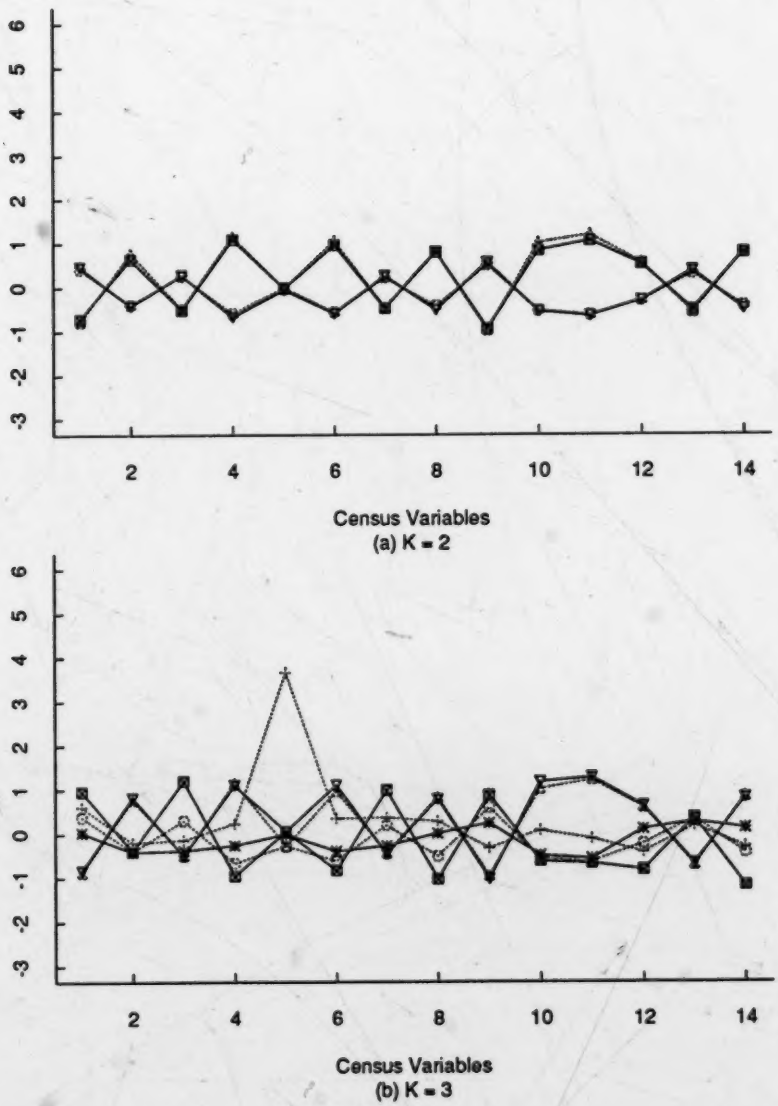


Figure 4.7: Representative Curves based on ($K = 2$ and 3) principal point estimates in Example 3 (Boston Housing Data). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines).

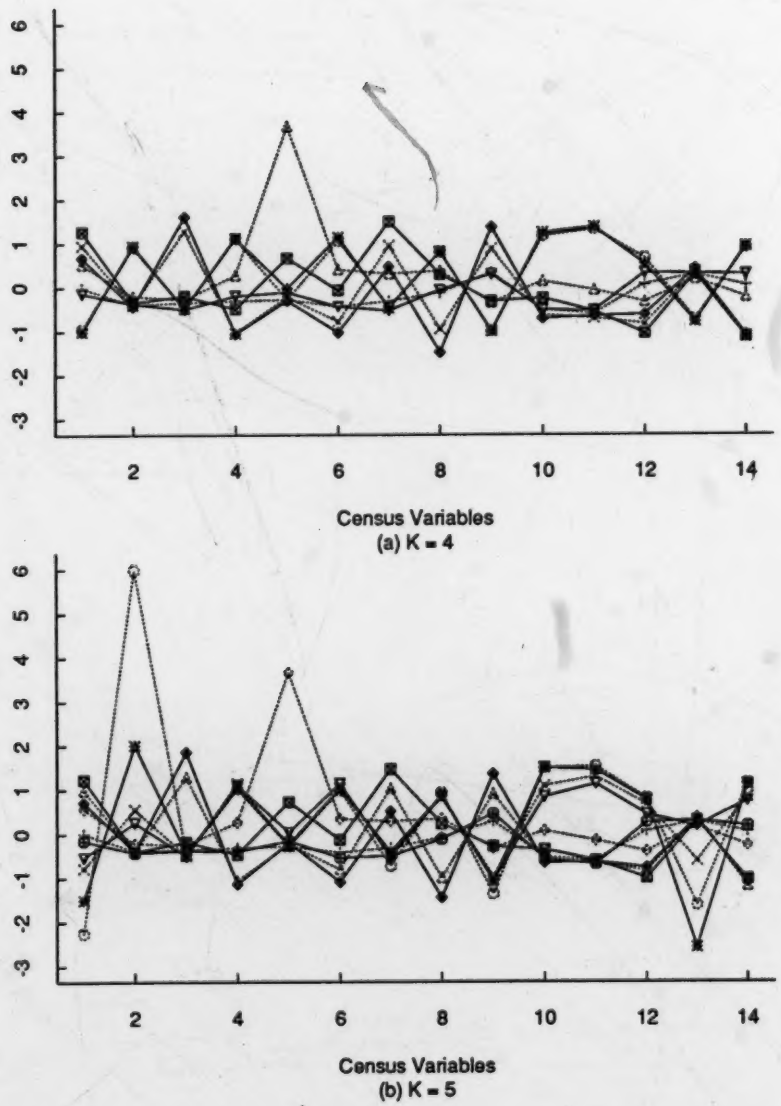


Figure 4.8: Representative Curves based on ($K = 4$ and 5) principal point estimates in Example 3 (Boston Housing Data). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines).

Table 4.4: Results for Seals Data.

$n = 3000$ and $p = 6$.

R_{opt} is the estimate of the risk function evaluated at $\tau = \tau_{PP}$.

R_0 is the estimate of the risk function evaluated at $\tau = 0$.

R^*_{opt} is the 30–fold cross–validated estimate of the risk function at $\tau = \tau_{PP}$.

R^*_0 is the 30–fold cross–validated estimate of the risk function at $\tau = 0$.

K	τ_{PP}	R_{opt}/R_0	R^*_{opt}/R^*_0
2	0.56711	0.97185	0.99654
3	0.53318	0.94645	0.98940
4	0.56919	0.89210	0.98802
5	0.81645	0.67113	0.90699
7	0.78756	0.62041	0.83873
8	0.79528	0.67056	0.83699
10	0.68459	0.72129	0.88895

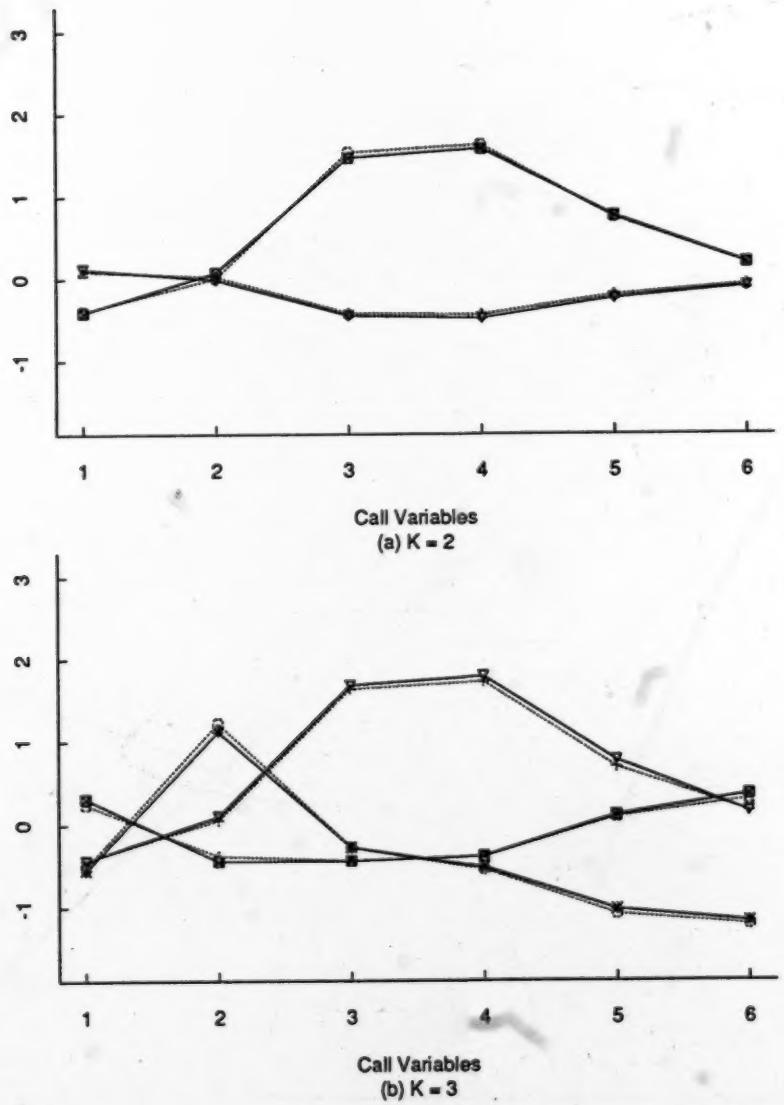


Figure 4.9: Representative Curves based on ($K = 2$ and 3) principal point estimates in Example 4 (Seals data). The curves based on principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines).

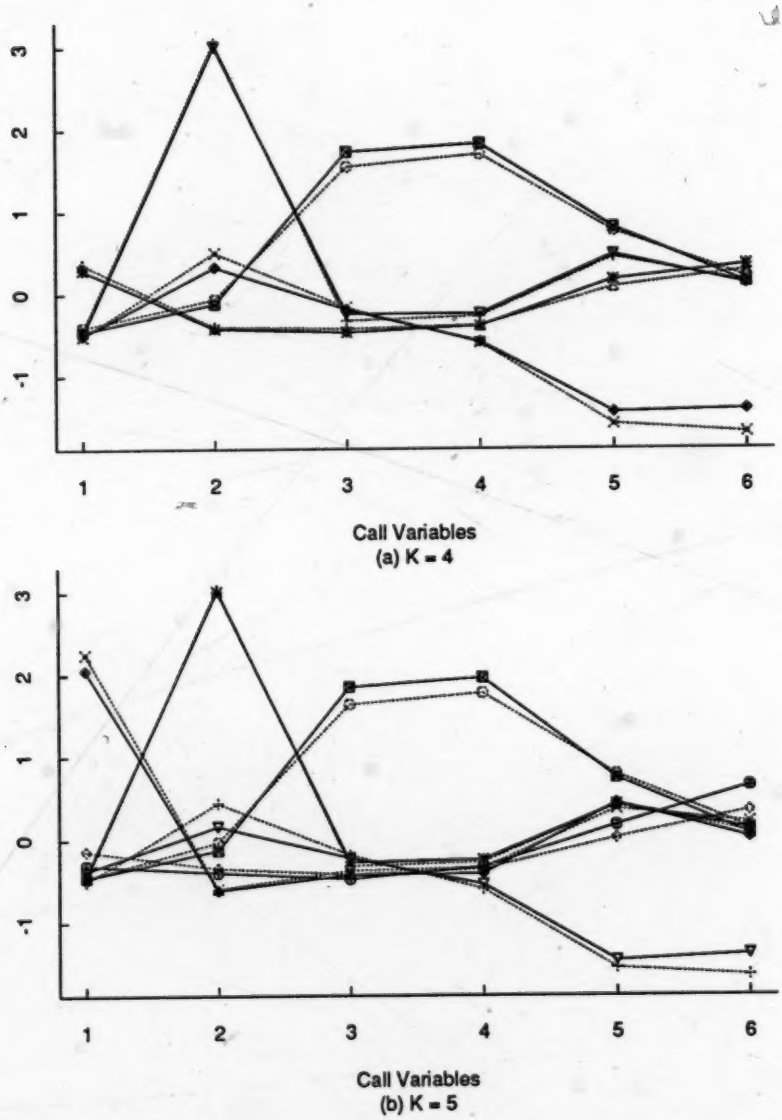


Figure 4.10: Representative Curves based on principal point estimates in Example 4 (Seals data). The curves based on ($K = 4$ and 5) principal points when $\tau = 0$ (dash lines) and principal points when $\tau = \tau_{PP}$ (solid lines).

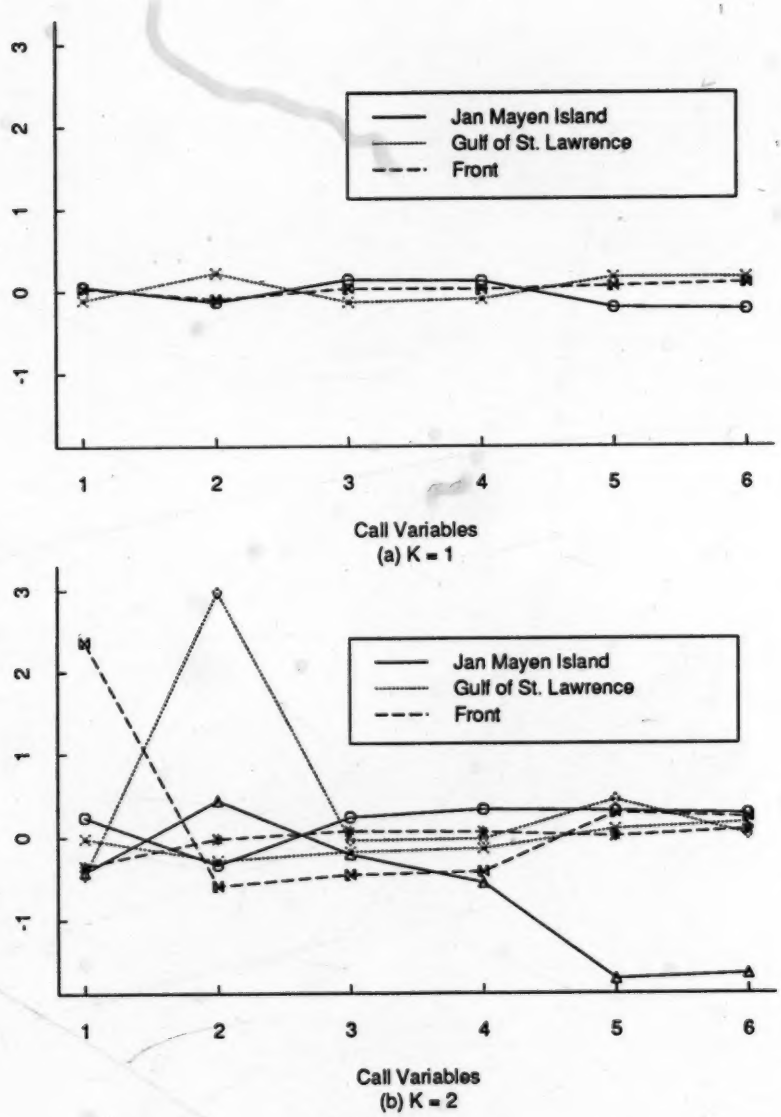


Figure 4.11: Representative Curves for each of the herds based on ($K = 1$ and 2) principal point estimates in Example 4 (Seals data).

Bibliography

[1] Belsely, D. E., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics.* New York: John Wiley.

[2] Benvensite, A., Métivier, M. and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations.* New York: Springer–Verlag.

[3] Breiman, L. and Friedman, J. H. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association.* **80**, 580–619.

[4] Celeux, G. and Govaert, G. (1992). A Classification EM Algorithm for Clustering and Two Stochastic Versions. *Computational Statistics and Data Analysis.* **14**, 315–332.

[5] Cochran, R. G. (1977). *Sampling Techniques.* New York: Wiley.

[6] Cohn, D., Riskin, E. and Ladner, R. (1994). Theory and Practice of Vector Quantizater on Small Training Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **16**, 54–65.

[7] Cox, D. R. (1957). Note on Grouping. *Journal of the American Statistical Association.* **52**, 543–547.

[8] Dalenius, T. (1950). The Problem of Optimum Stratification. *Skand. Aktuar.* **33**, 203–213.

[9] Dalenius, T. and Gurney, M. (1951). The Problem of Optimum Stratification II. *Skand. Atkuar.* **34**, 133–148.

[10] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, **39**, 1–38.

[11] Flury, B. D. (1990). Principal Points. *Biometrika.* **77**, 33–41.

[12] Flury, B. D. (1993). Estimation of Principal Points. *Applied Statistics.* **42**, 139–151.

[13] Flury, B. D. and Tarpey, T. (1993). Representing a Large Collection of Curves: A Case for Principal Points. *American Statistics.* **47**, 304–306.

[14] Gersho, R. and Gray, R. M. (1992). *Vector Quantization and Signal Compression.* Boston: Kluwer Academic Publishers.

[15] Harrison, D. and Rubinfeld, D. L. (1978). Hedonic Housing Prices and Demand for Clean Air. *Journal of Environmental Economics and Management.* **5**, 81–102.

[16] Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K–Means Clustering Algorithm. *Applied Statistics.* **28**, 100–108.

[17] Hooper, P. (1996a). Nonparametric Piecewise Linear Classification Rules. Submitted.

[18] Hooper, P. (1996b). Flexible Regression on Multiple Predictors by Randomized Exponential Smoothing. Submitted.

[19] Johns, M. C. and Rice, J. A. (1992). Displaying the Important Features of Large Collections of Similar Curves. *The American Statistician.* **46**, 140–145.

[20] Kohonen, T. (1995). *Self-Organizing Maps.* New York: Springer.

[21] MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1**, 281–297.

[22] Moody, J. E. and Darken, C. (1989). Fast Learning in Networks of Locally-tuned Processing Units. *Neural Computation.* **1**, 281–294.

[23] Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *Annals of Mathematical Statistics.* **22**, 103–112.

[24] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* New York: Chapman and Hall.

[25] Stone, M. (1974). Cross–Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Association*, Series B. **36**, 111–133.

[26] Tarpey, T. (1994). Two Principal Points of Symmetric, Strongly Unimodal Distributions. *Statistics and Probability Letters.* **20**, 253–257.

[27] Tarpey, T., Li, L., and Flury, B. D. (1995). Principal Points and Self-Consistent Points of Elliptical Distributions. *Annals of Statistics.* **23**, 103–112.